

On the analysis of inexact augmented Lagrangian schemes for misspecified conic convex programs*

H. AHMADI[†], N. S. AYBAT[†], AND U. V. SHANBHAG[†]

Abstract. We consider a misspecified optimization problem that requires minimizing a convex function $f(x; \theta^*)$ in x over a conic constraint set represented by $h(x; \theta^*) \in \mathcal{K}$, where θ^* is an unknown (or misspecified) vector of parameters, \mathcal{K} is a proper cone, and h is affine in x . Suppose θ^* is not available; but, it can be learnt by a separate process that generates a sequence of estimators θ_k , each of which is an increasingly accurate approximation of θ^* . We develop a first-order inexact augmented Lagrangian (AL) scheme for computing x^* while simultaneously learning θ^* . In particular, we derive rate statements for such schemes when the penalty parameter sequence is either constant or increasing, and derive bounds on the overall complexity in terms of proximal-gradient steps when AL subproblems are solved via an accelerated proximal-gradient scheme. Numerical results for a portfolio optimization problem with a misspecified covariance matrix suggest that these schemes perform well in practice. In particular, we note that naive sequential schemes for contending with misspecified optimization problems may perform poorly in practice.

Key words. augmented Lagrangian, misspecification, conic programming, learning, first-order algorithms

1. Introduction. Consider an optimization problem in n -dimensional Euclidean space defined as follows:

$$(\mathcal{C}(\theta^*)) : \quad \mathcal{X}^*(\theta^*) \triangleq \operatorname{argmin}_{x \in X \cap \mathcal{H}(\theta^*)} f(x; \theta^*), \quad (1.1)$$

where $\theta^* \in \mathbb{R}^d$ denotes the parametrization of the objective and constraints. Traditionally, optimization research has considered settings where θ^* is available a priori, a singular exception being robust optimization approaches.

Robust optimization: For instance, when θ^* is unavailable and $\mathcal{H}(\theta^*) = \mathbb{R}^n$, but one has access to an uncertainty set \mathcal{T} corresponding to θ , then robust optimization approaches minimize the worst-case value that $f(x, \theta)$ assumes on the set \mathcal{T} , as captured by the following formulation:

$$\min_{x \in X} \max_{\theta \in \mathcal{T}} f(x; \theta). \quad (1.2)$$

Robust optimization has proved to be an enormously useful technique in the resolution of problems in design, control, and optimization (See [6]). In this paper, motivated by the increasing accessibility to data, we consider an alternate approach in which θ has a nominal or true value θ^* obtainable by solving a suitably defined learning problem:

$$(\mathcal{E}) : \quad \min_{\theta \in \Theta} \ell(\theta). \quad (1.3)$$

Such problems routinely arise when θ^* is idiosyncratic to the problem and may be learnt by the aggregation of data; instances arise when attempting to learn covariance matrices associated with a collection of stocks, efficiency parameters associated with machines on a supply line, or demand parameters associated with a supply chain. A

*A preliminary version of this work will appear in the Proceedings of the American Control Conference [1]. This work has been partially funded by NSF Grant CMMI-1400217 (Ahmadi, Aybat, and Shanbhag) and NSF Grant CMMI -1246887 (CAREER) (Shanbhag).

[†]Department of Industrial and Manufacturing Engineering, the Pennsylvania State University, University Park, PA-16802. (ahmadi.hesam@gmail.com, nsa10@psu.edu, udaybag@psu.edu).

natural approach in this case is to first estimate θ^* with high accuracy and then solve the parametrized problem. Yet, in many instances, this *sequential* approach cannot be adopted for at least two reasons:

- (i) The learning problem can be large, *precluding* a highly accurate a priori parameter resolution in a reasonable time; hence, making it impractical to solve the original problem, i.e., the decision maker may have to wait for a long time during this learning process with *no* availability of an estimate solution to $\mathcal{C}(\theta^*)$;
- (ii) Unless the learning problem can be solved *exactly* in finite time, sequential schemes are not asymptotically convergent and can, at best, provide approximate solutions. Indeed, the lack of exactness arising from the error in learning cascades into the resolution of the subsequent optimization problem.

Accordingly, we consider the development of schemes that generate sequences $\{x_k\}, \{\theta_k\}$ such that

$$\|\theta_k - \theta^*\| \rightarrow 0, \quad d_{\mathcal{X}^*(\theta^*)}(x_k) \rightarrow 0 \quad \text{as } k \rightarrow \infty,$$

where θ^* is the unique solution of (\mathcal{E}) and $\mathcal{X}^*(\theta^*)$ is the optimal solution set for $\mathcal{C}(\theta^*)$; and $d_{\mathcal{X}}(x) \triangleq \min_{s \in \mathcal{X}} \|x - s\|$ denotes the distance function to a given closed convex set \mathcal{X} .

This work originates from prior work that has coupled stochastic approximation schemes with rate statements for stochastic optimization/variational inequality problems [14, 16] and misspecified stochastic Nash games [17]. Subsequently, these statements were refined and sharpened for deterministic optimization problems [3] and extended to the resolution of misspecified distributed stochastic optimization problems [18] and misspecified Markov Decision Processes (MDPs) [15]. This paper is inspired by the challenges arising from misspecification in the constraint set. Such concerns are addressed in [2, 14, 17] by considering the associated variational inequality problem. In sharp contrast, in this paper, we consider a more general form of misspecification in the constraints (namely convex conic). Rather than standard gradient-type approaches, we develop a misspecified analog of an augmented Lagrangian (AL) scheme for misspecified convex problems in which both the objective and the constraints are misspecified and the subproblems are solved with increasing exactness. Throughout, our focus will be on the problem $\mathcal{C}(\theta^*)$ when $\mathcal{H}(\theta^*) \triangleq \{x : h(x; \theta^*) \preceq_{\mathcal{K}} 0\}$, where $\preceq_{\mathcal{K}}$ denotes the partial order induced by the proper cone \mathcal{K} in \mathbb{R}^m , i.e., given $a, b \in \mathbb{R}^m$, $a \preceq_{\mathcal{K}} b$ implies $b - a \in \mathcal{K}$. Hence, $\mathcal{H}(\theta^*) \equiv \{x : -h(x; \theta^*) \in \mathcal{K}\}$.

Consequently, we may redefine the problem in parametric form, $\mathcal{C}(\theta)$, more explicitly as follows:

$$(\mathcal{C}(\theta)) : \quad \min_x \{f(x; \theta) : h(x; \theta) \preceq_{\mathcal{K}} 0, \quad x \in X\}, \quad (1.4)$$

where $f : \mathbb{R}^n \times \Theta \rightarrow \mathbb{R} \cup \{+\infty\}$, $h : \mathbb{R}^n \times \Theta \rightarrow \mathbb{R}^m$, \mathcal{K} is a proper cone in \mathbb{R}^m , i.e., closed, convex, pointed, with a nonempty interior, and $\theta \in \Theta \subseteq \mathbb{R}^d$ denotes the misspecified parameter. Throughout, we assume that $\mathcal{C}(\theta^*)$ has a finite optimal value, denoted by f^* , i.e., $f^* = f(x^*; \theta^*)$ for all $x^* \in X^*(\theta^*)$. Moreover, we also assume that the corresponding Lagrangian dual problem has a solution, denoted by λ^* , and there is *no* duality gap.

AL schemes are rooted in the seminal works by Hestenes [11] and Powell [24], and their relation to proximal-point methods is established by Rockafellar [25, 26]. More recently, there has been a renewed examination of such techniques, particularly in convex regimes, with an emphasis on deriving rate estimates, e.g., [4, 19, 21].

Next, we outline our main contributions:

- (i) In Section 3, we derive rate statements for dual suboptimality, primal infeasibility, and primal suboptimality for the prescribed coupled first-order scheme with a quantification of the impact of misspecification. In particular, in Section 3.1 we consider a setting with a constant penalty parameter, and in Section 3.2 we derive analogous rate statements in a setting where the penalty parameters is increased after each iteration
- (ii) An overall iteration complexity analysis of the scheme is provided in Section 4. In Section 4.1 we consider the constant penalty case and prove that at most $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-4})$ proximal-gradient steps are required to obtain an ϵ -feasible and ϵ -optimal solution without and with learning, respectively. Utilizing a suitably defined sequence of increasing penalty parameters, in Section 4.2, we note that this worst-case complexity reduces to $\mathcal{O}(\epsilon^{-1} \log(\epsilon^{-1}))$ regardless of whether a parallel learning process is employed. After having independently proven iteration complexity statements for a constant penalty AL scheme in our preliminary work [1], we became aware of related work [21] that considers inexact augmented Lagrangian schemes *without* learning. When θ^* is available, the complexity statements provided in this manuscript and in [1], while related to the statements provided in [21], are both novel and distinct.
- (iii) Finally, in Section 5, we demonstrate the utility of the prescribed scheme through a portfolio optimization problem with a misspecified covariance matrix where several aspects of the scheme become evident: (i) The misspecified variants of the augmented Lagrangian schemes perform well in practice; (ii) the complexity bound in the constant penalty case appears to be relatively loose; and (iii) naive sequential schemes can perform quite poorly in comparison with their proposed simultaneous counterparts.

Notation. Given a closed convex set $\mathcal{K} \subset \mathbb{R}^m$ and $y \in \mathbb{R}^m$, define $d_{\mathcal{K}}(y) \triangleq \min_{s \in \mathcal{K}} \|y - s\|$, and $d_{\mathcal{K}}^2(y) \triangleq (d_{\mathcal{K}}(y))^2$. We denote the Euclidean projection $\Pi_{\mathcal{K}}(y) \triangleq \operatorname{argmin}_{s \in \mathcal{K}} \|y - s\|$; hence, $d_{\mathcal{K}}(y) = \|y - \Pi_{\mathcal{K}}(y)\|$. Moreover, it is easy to verify that $d_{\mathcal{K}}^2(\cdot)$ is differentiable and its gradient $\nabla d_{\mathcal{K}}^2(y) = 2(y - \Pi_{\mathcal{K}}(y))$. Let $\mathcal{B}(\bar{y}, r) \triangleq \{y : \|y - \bar{y}\| \leq r\}$. Given a cone $\mathcal{K} \in \mathbb{R}^m$, let \mathcal{K}^* denote its dual cone, i.e., $\mathcal{K}^* = \{y' \in \mathbb{R}^m : \langle y', y \rangle \geq 0 \ \forall y \in \mathcal{K}\}$. Given $A \in \mathbb{R}^{m \times n}$, $\sigma_{\max}(A)$ denotes the largest singular value of A , and $\|A\| \triangleq \sigma_{\max}(A)$ denotes the spectral norm of A .

2. Preliminaries. Given $\theta \in \Theta$, the problem $\mathcal{C}(\theta)$ is equivalent to the following reformulated problem:

$$\min_{x, z} \{f(x; \theta) : h(x; \theta) + z = 0, \quad x \in X, \quad z \in \mathcal{K}\}. \quad (2.1)$$

Let $\lambda \in \mathbb{R}^m$ denote a dual variable corresponding to the equality constraints in (2.1). For any given $\rho > 0$, we denote the augmented Lagrangian function for (2.1) by $\mathcal{L}_{\rho}(x, \lambda; \theta)$ and define it as

$$\mathcal{L}_{\rho}(x, \lambda; \theta) \triangleq \min_{z \in \mathcal{K}} [f(x; \theta) + \lambda^{\top} (h(x; \theta) + z) + \frac{\rho}{2} \|h(x; \theta) + z\|^2],$$

where $\text{dom } \mathcal{L}_\rho = X \times \mathbb{R}^m \times \Theta$. Through a rearrangement of terms, it can be shown that

$$\begin{aligned}\mathcal{L}_\rho(x, \lambda; \theta) &= f(x; \theta) + \frac{\rho}{2} \min_{z \in \mathcal{K}} \left\| h(x; \theta) + z + \frac{\lambda}{\rho} \right\|^2 - \frac{\|\lambda\|^2}{2\rho} \\ &= f(x; \theta) + \frac{\rho}{2} d_{\mathcal{K}}^2 \left(- \left(h(x; \theta) + \frac{\lambda}{\rho} \right) \right) - \frac{\|\lambda\|^2}{2\rho}.\end{aligned}\quad (2.2)$$

Note that for any $\bar{y} \in \mathbb{R}^m$, we have

$$\Pi_{\mathcal{K}}(-\bar{y}) = \underset{y \in \mathcal{K}}{\operatorname{argmin}} \|y + \bar{y}\| = - \underset{\hat{y} \in -\mathcal{K}}{\operatorname{argmin}} \|\hat{y} - \bar{y}\| = -\Pi_{-\mathcal{K}}(\bar{y}); \quad (2.3)$$

and consequently, by invoking (2.3), we also have that

$$d_{\mathcal{K}}(-\bar{y}) = \| - \Pi_{\mathcal{K}}(-\bar{y}) - \bar{y} \| = \| \Pi_{-\mathcal{K}}(\bar{y}) - \bar{y} \| = d_{-\mathcal{K}}(\bar{y}). \quad (2.4)$$

It follows that by invoking (2.4), $\mathcal{L}_\rho(x, \lambda; \theta)$, given by (2.2), can be rewritten as

$$\mathcal{L}_\rho(x, \lambda; \theta) = f(x; \theta) + \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(x; \theta) + \frac{\lambda}{\rho} \right) - \frac{\|\lambda\|^2}{2\rho}. \quad (2.5)$$

Next, we derive the gradient $\nabla_\lambda \mathcal{L}_\rho(x, \lambda; \theta)$. To simplify notation, let $\psi(\lambda) \triangleq \frac{\lambda}{\rho} + h(x; \theta)$. Then, since $\nabla d_{\mathcal{K}}^2(x) = 2(x - \Pi_{\mathcal{K}}(x))$ for $x \in \mathbb{R}^m$, $\nabla_\lambda \mathcal{L}_\rho$ can be computed as follows:

$$\nabla_\lambda \mathcal{L}_\rho(x, \lambda; \theta) = h(x; \theta) - \Pi_{-\mathcal{K}} \left(\frac{\lambda}{\rho} + h(x; \theta) \right) = \Pi_{\mathcal{K}^*} \left(\frac{\lambda}{\rho} + h(x; \theta) \right) - \frac{\lambda}{\rho}, \quad (2.6)$$

where in the last equality, we use the property that $\bar{x} = \Pi_{-\mathcal{K}}(\bar{x}) + \Pi_{\mathcal{K}^*}(\bar{x})$ for all $\bar{x} \in \mathbb{R}^m$. Clearly, $\mathcal{L}_0(x, \lambda; \theta)$, i.e., $\mathcal{L}_\rho(x, \lambda; \theta)|_{\rho=0}$ is the Lagrangian function:

$$\mathcal{L}_0(x, \lambda; \theta) \triangleq \begin{cases} f(x; \theta) + \lambda^\top h(x; \theta), & \text{if } \lambda \in \mathcal{K}^* \\ -\infty, & \text{otherwise.} \end{cases} \quad (2.7)$$

For $\rho \geq 0$, the *augmented* dual problem of $\mathcal{C}(\theta)$ is defined as

$$(D_\rho) : \quad \max_{\lambda \in \mathcal{K}^*} \left\{ g_\rho(\lambda; \theta) \triangleq \inf_{x \in X} \mathcal{L}_\rho(x, \lambda; \theta) \right\}.$$

Further, $X^*(\lambda; \theta)$ denotes the solution set of the Lagrangian problem:

$$X^*(\lambda; \theta) \triangleq \underset{x \in X}{\operatorname{argmin}} \mathcal{L}_0(x, \lambda; \theta).$$

Clearly, $\text{dom } g_\rho \subseteq \mathcal{K}^*$ for $\rho \geq 0$. Throughout the paper, we make the following additional assumptions:

Assumption 1.

- (i) Suppose $X \subseteq \mathbb{R}^n$ and Θ are convex compact sets. In addition, the function $f(x, \theta)$ is convex in $x \in X$ for all $\theta \in \Theta$ and Lipschitz continuous in $\theta \in \Theta$ uniformly for all $x \in X$ with constant L_f ; i.e., for all $x \in X$, $\|f(x; \theta_1) - f(x; \theta_2)\| \leq L_f \|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$.

- (ii) $h(x; \theta)$ is an affine map in x for every $\theta \in \Theta$, i.e., $h(x; \theta) = A(\theta)x + b(\theta)$ for some $A(\theta) \in \mathbb{R}^{m \times n}$ and $b(\theta) \in \mathbb{R}^m$. Suppose $A(\theta)$ and $b(\theta)$ are Lipschitz continuous in θ , i.e., there exist constants L_A, L_b such that for all $\theta_1, \theta_2 \in \Theta$,

$$\|A(\theta_1) - A(\theta_2)\| \leq L_A \|\theta_1 - \theta_2\|, \quad \|b(\theta_1) - b(\theta_2)\| \leq L_b \|\theta_1 - \theta_2\|.$$

- (iii) $X^*(\lambda; \theta)$ is pseudo-Lipschitz in θ uniformly for $\lambda \in \mathcal{K}^*$, i.e., there exists a constant κ_X such that for any $\theta_1, \theta_2 \in \Theta$, $X^*(\lambda; \theta_1) \subseteq X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)$ for all $\lambda \in \mathcal{K}^*$.

Remark 2.1. Given $x \in X$, by Assumption 1 (ii), for all $\theta, \theta' \in \Theta$,

$$\|h(x; \theta) - h(x; \theta')\| \leq \|A(\theta) - A(\theta')\| \|x\| + \|b(\theta) - b(\theta')\| \leq (L_A \|x\| + L_b) \|\theta - \theta'\|.$$

Since X is assumed to be compact, there exists a finite constant D_x such that $D_x \triangleq \max_{x \in X} \|x\|$. Hence, $h(x; \theta)$ is Lipschitz continuous in θ with constant $L_{h, \theta}$ where $L_{h, \theta} \triangleq (L_A D_x + L_b)$ uniformly for all $x \in X$. Moreover, given $\theta \in \Theta$,

$$\|h(x; \theta) - h(x'; \theta)\| \leq \|A(\theta)\| \|x - x'\|$$

for all $x, x' \in X$. Since $\|A(\theta)\| = \sigma_{\max}(A(\theta))$ is continuous in θ and Θ is compact, $L_{h, x} \triangleq \max_{\theta \in \Theta} \|A(\theta)\|$ exists; therefore, $h(x; \theta)$ is Lipschitz continuous in x , uniformly for all $\theta \in \Theta$, with constant $L_{h, x}$. Clearly, for all $\theta \in \Theta$, $h(\mathcal{B}(\mathbf{0}, 1); \theta) \subseteq \mathcal{B}(b(\theta), \|A(\theta)\|)$; hence, $h(\mathcal{B}(\mathbf{0}, 1); \theta) \subseteq b(\theta) + L_{h, x} \mathcal{B}(\mathbf{0}, 1)$. Rather than focusing on the nature of the algorithm employed for resolving the learning problem, we assume that the adopted scheme produces a sequence that converges to the optimal solution θ^* at a non-asymptotic **linear** rate (Assumption 2).

Assumption 2. There exists a learning scheme that generates a sequence $\{\theta_k\}$ such that $\theta_k \rightarrow \theta^*$ at a linear rate as $k \rightarrow \infty$, i.e., there exists a constant $\tau \in (0, 1)$ such that for all $k \geq 0$ and $\theta_0 \in \Theta$, one has $\|\theta_k - \theta^*\| \leq \tau^k \|\theta_0 - \theta^*\|$. In addition, at iteration k of the optimization problem \mathcal{C} , only $\theta_1, \dots, \theta_k$ are revealed. Lemma 2.1 provides various properties of the gradient of the dual function $\nabla_{\lambda} g_{\rho}$ and will be used in our analysis. Its proof may be found in [26] and is omitted.

LEMMA 2.1. Suppose Assumption 1 holds.

- (i) For any $\rho > 0$ and $\theta \in \Theta$, the dual function $g_{\rho}(\lambda; \theta)$ is everywhere finite, continuously differentiable concave function over \mathbb{R}^m ; more precisely, $g_{\rho}(\lambda; \theta) = \max_{w \in \mathbb{R}^m} \{g_0(w; \theta) - \frac{1}{2\rho} \|w - \lambda\|^2\}$, i.e., $g_{\rho}(\cdot, \theta)$ is the Moreau envelope of $g_0(\cdot, \theta)$ for all $\theta \in \Theta$. Therefore, $\nabla_{\lambda} g_{\rho}(\lambda; \theta)$ is Lipschitz continuous in λ with constant $\frac{1}{\rho}$ for all θ in Θ .
- (ii) For any given $\lambda \in \mathcal{K}^*$ and $\theta \in \Theta$, $\nabla_{\lambda} g_{\rho}$ can be computed as $\nabla_{\lambda} g_{\rho}(\lambda; \theta) = \nabla_{\lambda} \mathcal{L}_{\rho}(x^*(\lambda; \theta), \lambda; \theta)$, where $x^*(\lambda; \theta) \in \operatorname{argmin}_{x \in X} \mathcal{L}_{\rho}(x, \lambda; \theta)$.
- (iii) Given $\lambda \in \mathcal{K}^*$, $\theta \in \Theta$, let $\tilde{x}(\lambda; \theta)$ be an inexact solution to $\min_{x \in X} \mathcal{L}_{\rho}(x, \lambda; \theta)$ with accuracy α , i.e., $\tilde{x}(\lambda; \theta) \in X$ satisfies $\mathcal{L}_{\rho}(\tilde{x}(\lambda; \theta), \lambda; \theta) \leq g_{\rho}(\lambda; \theta) + \alpha$, then

$$\|\nabla_{\lambda} \mathcal{L}_{\rho}(\tilde{x}(\lambda; \theta), \lambda; \theta) - \nabla_{\lambda} g_{\rho}(\lambda; \theta)\|^2 \leq \frac{2\alpha}{\rho}.$$

Next, we examine the Lipschitz continuity of $\nabla_{\lambda} g_{\rho}(\lambda; \theta)$ in $\theta \in \Theta$. Recall that

$$g_{\rho}(\lambda; \theta) = \max_w \left[g_0(w; \theta) - \frac{1}{2\rho} \|w - \lambda\|^2 \right].$$

By properties of the Moreau envelope [12], it follows that

$$\nabla_\lambda g_\rho(\lambda; \theta) = \frac{1}{\rho}(\pi_\rho(\lambda; \theta) - \lambda), \quad (2.8)$$

where $\pi_\rho(\cdot; \theta)$, the Moreau map of $g_\rho(\cdot; \theta)$, is defined as:

$$\pi_\rho(\lambda; \theta) \triangleq \operatorname{argmax}_w g_0(w; \theta) - \frac{1}{2\rho} \|w - \lambda\|^2. \quad (2.9)$$

Therefore, it suffices to show the Lipschitz continuity of $\pi_\rho(\lambda; \theta)$ in θ . We begin with an intermediate Lemma that proves the Lipschitz continuity of π_ρ under a suitable pseudo-Lipschitzian property on the *supdifferential* of $g_0(\cdot; \theta)$, i.e.,

$$\partial_w g_0(w; \theta) \triangleq \{s \in \mathbb{R}^m : g_0(w; \theta) + s^\top(w' - w) \geq g_0(w; \theta) \quad \forall w' \in \mathcal{K}^*\}.$$

LEMMA 2.2. *Suppose there exists a κ such that $g_0(w; \theta)$ satisfies the following for all w and for all $\theta_1, \theta_2 \in \Theta$:*

$$\partial_w g_0(w; \theta_1) \subseteq \partial_w g_0(w; \theta_2) + \kappa \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1), \quad (2.10)$$

where $\mathcal{B}(\mathbf{0}, 1) = \{z : \|z\| \leq 1\}$. Then the following holds:

$$\|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\| \leq \kappa \rho \|\theta_1 - \theta_2\| \text{ for all } \theta_1, \theta_2 \in \Theta. \quad (2.11)$$

Proof. Recall that $\pi_\rho(\lambda; \theta)$ is defined by (2.9), implying the following:

$$0 \in \partial_w g_0(\pi_\rho(\lambda; \theta_1); \theta_1) + \frac{1}{\rho}(\pi_\rho(\lambda; \theta_1) - \lambda) \implies \frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_1)) \in \partial_w g_0(\pi_\rho(\lambda; \theta_1); \theta_1).$$

Similarly, we have that

$$\frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_2)) \in \partial_w g_0(\pi_\rho(\lambda; \theta_2); \theta_2).$$

Consequently, from the assumption in (2.10), there exists a vector $\xi \in \kappa \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)$ such that

$$\frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_1)) \in \partial_w g_0(\pi_\rho(\lambda; \theta_1); \theta_1) \subseteq \partial_w g_0(\pi_\rho(\lambda; \theta_1); \theta_2) + \xi.$$

Therefore, we have that

$$\frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_1)) - \xi \in \partial_w g_0(\pi_\rho(\lambda; \theta_1); \theta_2).$$

By the monotonicity of the map $\partial_w g_0(\cdot; \theta)$ for every $\theta \in \Theta$ and for nonnegative ρ , we have that

$$\begin{aligned} 0 &\leq \left(\frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_1)) - \xi - \frac{1}{\rho}(\lambda - \pi_\rho(\lambda; \theta_2)) \right)^\top (\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)) \\ &= \frac{1}{\rho}(\pi_\rho(\lambda; \theta_2) - \pi_\rho(\lambda; \theta_1) - \rho\xi)^\top (\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)) \\ &= -\frac{1}{\rho} \|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\|^2 - \xi^\top (\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)). \end{aligned}$$

By rearranging the terms, we obtain the following inequality:

$$\begin{aligned} \frac{1}{\rho} \|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\|^2 &\leq \|\xi\| \|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\| \\ \implies \frac{1}{\rho} \|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\| &\leq \|\xi\|. \end{aligned}$$

Moreover, $\xi \in \kappa \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)$ implies that $\|\xi\| \leq \kappa \|\theta_1 - \theta_2\|$, which leads to (2.11). \square

The conditions under which the above results hold are not ideal in that they are assumed on $g(\lambda; \theta)$. However, as the next result shows, by assuming a suitable pseudo-Lipschitzian property on $X^*(\lambda; \theta)$ in θ uniformly in λ as in Assumption 1(iii), we obtain the required property.

LEMMA 2.3. *Under Assumption 1, $\partial_\lambda g_0(\lambda; \theta)$ is pseudo-Lipschitz in θ uniformly in λ with constant $\kappa \triangleq L_{h,\theta} + \kappa_X L_{h,x}$, i.e., $\partial_\lambda g_0(\lambda; \theta_1) \subseteq \partial_\lambda g_0(\lambda; \theta_2) + \kappa \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)$ for $\theta_1, \theta_2 \in \Theta$.*

Proof. Recall that $g_0(\lambda; \theta)$ is defined as $g_0(\lambda; \theta) \triangleq \min_{x \in X} \mathcal{L}_0(x, \lambda; \theta)$ where $\mathcal{L}_0(\cdot, \cdot; \theta)$ is defined in (2.7). Then for any $\lambda \in \mathcal{K}^*$ and $\theta \in \Theta$, Danskin's theorem implies that

$$\partial_\lambda g_0(\lambda; \theta) = \text{conv} \{h(x; \theta) : x \in X^*(\lambda; \theta)\}.$$

As a consequence, one may note that $\partial_\lambda g_0(\lambda; \theta)$ is given by the following:

$$\partial_\lambda g_0(\lambda; \theta) = \text{conv} \{h(X^*(\lambda; \theta); \theta)\} = h(X^*(\lambda; \theta); \theta), \quad (2.12)$$

since $h(x; \theta)$ is an affine map in x for every $\theta \in \Theta$ and $X^*(\lambda; \theta)$ is a convex set for any θ and λ , it follows that the image $h(X^*(\lambda; \theta); \theta)$ is also a convex set. By Assumption 1(iii), there exists a κ_X such that

$$X^*(\lambda; \theta_1) \subseteq X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1). \quad (2.13)$$

Since $h(\cdot; \theta)$ is an affine map for every $\theta \in \Theta$, from (2.13), it follows that

$$h(X^*(\lambda; \theta_1); \theta_1) \subseteq h(X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1); \theta_1). \quad (2.14)$$

We define $\bar{h}(x; \theta) \triangleq h(x; \theta) - b(\theta) = A(\theta)x$, as the linear part of $h(x; \theta)$. Then, the image of the Minkowski sum of sets, $\bar{h}(X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1); \theta_1)$, can be written as follows:

$$\begin{aligned} & \bar{h}(X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1); \theta_1) \\ &= \{\bar{h}(x + y; \theta_1) : x \in X^*(\lambda; \theta_2), y \in \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)\} \\ &= \{\bar{h}(x; \theta_1) + \bar{h}(y; \theta_1) : x \in X^*(\lambda; \theta_2), y \in \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1)\} \\ &= \{\bar{h}(x; \theta_1) + z : x \in X^*(\lambda; \theta_2), z \in \kappa_X \|\theta_1 - \theta_2\| \bar{h}(\mathcal{B}(\mathbf{0}, 1); \theta_1)\} \\ &\subseteq \bar{h}(X^*(\lambda; \theta_2); \theta_1) + \kappa_X \|A(\theta_1)\| \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1) \\ &\subseteq \bar{h}(X^*(\lambda; \theta_2); \theta_2) + L_A D_x \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1) + \kappa_X L_{h,x} \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1). \end{aligned} \quad (2.15)$$

where (2.16) follows from Assumption 1(ii) and the definitions of \bar{h} and $L_{h,x}$. By adding $b(\theta_1)$ to the both side of above inclusion, and using $h(x; \theta) = \bar{h}(x; \theta) + b(\theta)$, we obtain

$$\begin{aligned} & h(X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1); \theta_1) \\ &\subseteq h(X^*(\lambda; \theta_2); \theta_2) + (L_A D_x + \kappa_X L_{h,x}) \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1) + b(\theta_1) - b(\theta_2). \end{aligned} \quad (2.17)$$

Since $b(\theta)$ is Lipschitz continuous in θ with constant L_b and $L_{h,\theta} = L_A D_x + L_b$, we can rewrite (2.17) as follows:

$$\begin{aligned} & h(X^*(\lambda; \theta_2) + \kappa_X \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1); \theta_1) \\ &\subseteq h(X^*(\lambda; \theta_2); \theta_2) + (L_{h,\theta} + \kappa_X L_{h,x}) \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1). \end{aligned} \quad (2.18)$$

According to (2.12), $h(X^*(\lambda; \theta_i); \theta_i) = \partial_\lambda g_0(\lambda; \theta_i)$ for $i = 1, 2$; hence, from (2.14) and (2.18), it follows that

$$\partial_\lambda g_0(\lambda; \theta_1) \subseteq \partial_\lambda g_0(\lambda; \theta_2) + (L_{h,\theta} + \kappa_X L_{h,x}) \|\theta_1 - \theta_2\| \mathcal{B}(\mathbf{0}, 1). \quad (2.19)$$

□

We are ready to prove our main Lipschitzian property for $\pi(\lambda; \theta)$ and $\nabla_\lambda g_\rho(\lambda; \theta)$.

PROPOSITION 2.4. *Suppose Assumption 1 holds and let $\kappa \triangleq L_{h,\theta} + \kappa_X L_{h,x}$. Then, we have the following:*

- (i) $\|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\| \leq \kappa \rho \|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$.
- (ii) $\nabla_\lambda g_\rho(\lambda; \theta)$ is Lipschitz continuous in θ over Θ uniformly in $\lambda \in \mathcal{K}^*$ with constant κ .

Proof. The desired results follow from our previous observations:

- (i) This follows by invoking Lemma 2.2 and Lemma 2.3.
- (ii) $\nabla_\lambda g_\rho(\lambda; \theta)$ can be explicitly stated in terms of $\pi_\rho(\lambda; \theta)$ as given by (2.8). Consequently, for all θ_1, θ_2 in Θ , we have that $\|\nabla_\lambda g_\rho(\lambda; \theta_1) - \nabla_\lambda g_\rho(\lambda; \theta_2)\| = \frac{1}{\rho} \|\pi_\rho(\lambda; \theta_1) - \pi_\rho(\lambda; \theta_2)\| \leq \kappa \|\theta_1 - \theta_2\|$.

□

Remark: Since Assumption 1 necessitates that the solution set $X^*(\theta; \lambda)$ is Lipschitz continuous in θ uniformly over $\lambda \in \mathcal{K}^*$, we briefly comment on the conditions under which this indeed holds. For every $\theta \in \Theta$, if $f(x; \theta)$ is a differentiable convex function in x , and $h(x; \theta)$ is an affine function in x , then $X^*(\lambda; \theta)$ is the solution set of the variational inequality problem $\text{VI}(X, \nabla_x \mathcal{L}_0(\cdot, \lambda; \theta))$ for all $\lambda \in \mathcal{K}^*$ and $\theta \in \Theta$. We consider two sets of problem classes in providing conditions under which the associated solution sets admit pseudo-Lipschitzian properties:

(i). *Parametrized convex quadratic programming:* If for every $\theta \in \Theta$, $f(x; \theta)$ is a convex quadratic function in x , i.e., $f(x; \theta) = \frac{1}{2} x^\top Q(\theta) x + r(\theta)^\top x$, and $X(\theta)$ is a polyhedral set defined as

$$X(\theta) \triangleq \{x : A(\theta)x - b(\theta) \geq 0, x \geq 0\},$$

then a primal-dual optimal pair $(x(\theta), \lambda(\theta))$ to $\min_x \{f(x; \theta) : x \in X(\theta)\}$ for any given $\theta \in \Theta$ is also a solution to the *linear complementarity problem* $\text{LCP}(q(\theta), M(\theta))$ where

$$q(\theta) \triangleq \begin{pmatrix} r(\theta) \\ -b(\theta) \end{pmatrix} \text{ and } M(\theta) \triangleq \begin{pmatrix} Q(\theta) & -A(\theta) \\ A(\theta) & \mathbf{0} \end{pmatrix}.$$

Recall that z is a solution to $\text{LCP}(q, M)$ if $0 \leq z \perp Mz + q \geq 0$, where for $u, v \in \mathbb{R}^n$, $u \perp v$ denotes that $[u]_i [v]_i = 0$ for $i = 1, \dots, n$. Consider a $\tilde{\theta} \in \Theta$ and let $q(\tilde{\theta}) \in \text{int}(S(\tilde{\theta}))$ where $S(\tilde{\theta})$ denotes the solution set of $\text{LCP}(\mathbf{0}, M(\tilde{\theta}))$. By [8, Theorem 7.5.1], there exist positive parameters ϵ and κ_S such that for all $\theta \in \hat{\Theta}(\tilde{\theta})$,

$$S(q(\theta), M(\theta)) \subseteq S(q(\tilde{\theta}), M(\tilde{\theta})) + \kappa_S (\|q(\theta) - q(\tilde{\theta})\| + \|M(\theta) - M(\tilde{\theta})\|) \mathcal{B}(\mathbf{0}, 1),$$

where

$$\hat{\Theta}(\tilde{\theta}) \triangleq \left\{ \theta \in \Theta : \|q(\theta) - q(\tilde{\theta})\| + \|M(\theta) - M(\tilde{\theta})\| \leq \epsilon \right\}.$$

This is a local Lipschitzian requirement and under a suitable compactness assumption on Θ , it may be globalized.

(ii). *Parametrized convex programming*: More generally, when $f(x; \theta)$ is a nonlinear convex function of x for any given $\theta \in \Theta$, one needs to appeal to more powerful stability statements in the context of parametric variational inequality problems. Suppose $\mathcal{B}(H; \epsilon, S)$ denotes an ϵ -neighborhood of the function H comprising of all continuous functions G such that

$$\|G - H\|_S \triangleq \sup_{y \in S} \|G(y) - H(y)\| < \epsilon.$$

Then, given $\lambda \in \mathcal{K}^*$ and $\theta \in \Theta$, we qualify the associated $\text{VI}(X, \nabla_x \mathcal{L}(\cdot, \lambda; \theta))$ as *semi-stable* if there exist scalars $c, \epsilon > 0$ such that

$$X^*(\lambda; \hat{\theta}) \subseteq X^*(\lambda; \theta) + c \sup_{x \in X} \|\nabla_x \mathcal{L}(x, \lambda; \hat{\theta}) - \nabla_x \mathcal{L}(x, \lambda; \theta)\| \mathcal{B}(\mathbf{0}, 1),$$

for every $\hat{\theta} \in \Theta$ satisfying $\nabla_x \mathcal{L}(\cdot, \lambda; \hat{\theta}) \in \mathcal{B}(\nabla_x \mathcal{L}(\cdot, \lambda; \theta); \epsilon, X)$. In fact, a necessary and sufficient condition for semi-stability of $\text{VI}(X, F)$ is the following [10, Prop. 5.5.5]: There exists two positive scalars c and ϵ , such that for all $q \in \mathbb{R}^n$,

$$\|q\| < \epsilon \implies \text{SOL}(X, q + F) \subseteq \text{SOL}(X, F) + \mathcal{B}(\mathbf{0}, c\|q\|).$$

As part of future research, we intend to refine these statements so that they are customized to the regime of variational inequality problems with parametrized maps $\text{VI}(X, F(\cdot; \theta))$ to provide conditions on $f(x; \theta)$ that ensure the required pseudo-Lipschitzian properties on the solution sets $X^*(\lambda; \theta)$.

Algorithm 1 ALM – Misspecified inexact augmented Lagrangian scheme

Given $\lambda_0 \in \mathcal{K}^*$ and $x_0 \in X$, let $\{\rho_k\}$, $\{\alpha_k\}$ and $\{\theta_k\}$ be given sequences.

Then, for all $k \geq 0$, update:

1. Compute $x_{k+1} \in X$ such that $\mathcal{L}_{\rho_k}(x_{k+1}, \lambda_k; \theta_k) \leq g_{\rho_k}(\lambda_k; \theta_k) + \alpha_k$;
 2. $\lambda_{k+1} \leftarrow \lambda_k + \rho_k \nabla_{\lambda} \mathcal{L}_{\rho_k}(x_{k+1}, \lambda_k, \theta_k)$;
 3. $k \leftarrow k + 1$;
-

3. Misspecified inexact augmented Lagrangian scheme. In this section, we introduce the misspecified variant of the inexact augmented Lagrangian scheme, displayed in Algorithm 1. From Step 2 of Algorithm 1 and (2.6), it follows that for $k \geq 0$ we have

$$\lambda_{k+1} = \lambda_k + \rho_k \nabla_{\lambda} \mathcal{L}_{\rho_k}(x_{k+1}, \lambda_k, \theta_k) \tag{3.1a}$$

$$\begin{aligned} &= \lambda_k + \rho_k \left(h(x_{k+1}; \theta_k) - \Pi_{-\mathcal{K}} \left(\frac{\lambda_k}{\rho_k} + h(x_{k+1}; \theta_k) \right) \right) \\ &= \lambda_k + \rho_k \left(\Pi_{\mathcal{K}^*} \left(\frac{\lambda_k}{\rho_k} + h(x_{k+1}; \theta_k) \right) - \frac{\lambda_k}{\rho_k} \right) \\ &= \Pi_{\mathcal{K}^*} (\lambda_k + \rho_k h(x_{k+1}; \theta_k)). \end{aligned} \tag{3.1b}$$

Hence, $\{\lambda_k\} \subseteq \mathcal{K}^*$ for all $k \geq 0$. Notably, if $\theta_k = \theta^*$ for all $k \geq 0$, this reduces to the traditional version considered in [26]. The remainder of this section comprises of two subsections. We analyze the rate of convergence of Algorithm 1 for a constant penalty parameter in Section 3.1 and proceed to examine the increasing penalty parameter regime in Section 3.2.

3.1. Convergence analysis for constant penalty sequence $\rho_k = \rho > 0$. In this section, we study the convergence behavior of Algorithm 1 when $\rho_k = \rho$ for $k \geq 0$. We make the following assumption on the sequence of inexactness $\{\alpha_k\}$.

Assumption 3. *The inexactness sequence $\{\alpha_k\}$ satisfies $\sum_{k=0}^{\infty} \sqrt{\alpha_k} < \infty$. Define \bar{x}_k and $\bar{\lambda}_k$ as $\bar{x}_k \triangleq \frac{1}{k} \sum_{i=1}^k x_i$ and $\bar{\lambda}_k \triangleq \frac{1}{k} \sum_{i=1}^k \lambda_i$ for $k \geq 1$. Under Assumption 3, we show the following:*

- (i) $0 \leq f^* - g_\rho(\bar{\lambda}_k) \leq \mathcal{O}(1/k)$;
- (ii) $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \mathcal{O}(1/\sqrt{k})$,
- (iii) $-\mathcal{O}(1/\sqrt{k}) \leq f(\bar{x}_k; \theta^*) - f^* \leq \mathcal{O}(1/k)$.

and these statements are then utilized in deriving the overall computational complexity in Section 4.1.

After proving these bounds independently, we became aware of related recent work [21], where Algorithm 1 is considered with $\alpha_k = \alpha > 0$ for all $k \geq 0$ for some fixed $\alpha > 0$, and under the *perfect information* assumption, i.e., $\theta_k = \theta^*$ for all $k \geq 0$. In [21], it is shown that (i) $f^* - g_\rho(\bar{\lambda}_k) \leq \mathcal{O}(1/k) + \alpha$, (ii) $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \mathcal{O}(1/\sqrt{k})$, and (iii) $-\mathcal{O}(1/\sqrt{k}) \leq f(\bar{x}_k; \theta^*) - f^* \leq \mathcal{O}(1/k) + \alpha$. Therefore, according to [21], α should be fixed as a small constant as it appears in *both* primal and dual suboptimality bounds. Moreover, since α is fixed in [21], suboptimality of the iterate sequence may *stall* after certain iterations. In contrast, our method may start with large α_0 and gradually decrease it, ensuring *both* numerical stability and asymptotic convergence to optimality in contrast with [21] where the scheme provides approximate solutions at best.

We begin by showing that dual variables stay bounded by using a supporting Lemma whose proof follows from Lemma 2.1(i) and the properties of proximal maps (cf. [13]).

LEMMA 3.1. *Let $\pi_\rho(\lambda; \theta)$ be the proximal map of $g_0(\cdot; \theta)$ defined in (2.9). Then, for all $\lambda_1, \lambda_2 \in \mathcal{K}^*$ and $\theta \in \Theta$,*

$$\|\pi_\rho(\lambda_1; \theta) - \pi_\rho(\lambda_2; \theta)\|^2 + \|\pi_\rho^c(\lambda_1; \theta) - \pi_\rho^c(\lambda_2; \theta)\|^2 \leq \|\lambda_1 - \lambda_2\|^2,$$

where $\pi_\rho^c(\lambda; \theta) \triangleq \lambda - \pi_\rho(\lambda; \theta)$. Hence, π_ρ is nonexpansive in $\lambda \in \mathcal{K}^*$ for all $\theta \in \Theta$.

Proof. The proof is given in [29]. \square

Next, we prove that the sequence $\{\lambda_k\}$ stays bounded, mainly using the same proof technique in [25, 27] and combining it with Lipschitz continuity result in Proposition 2.4. the general case where penalty parameter sequence is allowed to change.

PROPOSITION 3.2 (Boundedness of $\{\lambda_k\}$). *Let Assumptions 1–3 hold, and λ^* be an arbitrary solution to the Lagrangian dual of $\mathcal{C}(\theta^*)$, i.e., $\lambda^* \in \operatorname{argmax}_\lambda g_0(\lambda; \theta^*)$. Then for all $k \geq 1$, $\|\lambda_k - \lambda^*\| \leq C_\lambda$, where C_λ is defined as follows:*

$$C_\lambda \triangleq \sqrt{2\rho} \sum_{i=0}^{\infty} \sqrt{\alpha_i} + \rho\kappa \frac{\|\theta_0 - \theta^*\|}{1 - \tau} + \|\lambda_0 - \lambda^*\|. \quad (3.2)$$

Proof. We begin by deriving a bound on $\|\lambda_{k+1} - \pi_\rho(\lambda_k; \theta_k)\|$. From (2.8) and the definition of λ_{k+1} , given in Step 2 of Algorithm 1, it follows that

$$\begin{aligned} \|\lambda_{k+1} - \pi_\rho(\lambda_k; \theta_k)\| &= \|\lambda_k + \rho \nabla_\lambda \mathcal{L}_\rho(x_{k+1}, \lambda_k; \theta_k) - \lambda_k - \rho \nabla_\lambda g_\rho(\lambda_k; \theta_k)\| \\ &= \rho \|\nabla_\lambda \mathcal{L}_\rho(x_{k+1}, \lambda_k; \theta_k) - \nabla_\lambda g_\rho(\lambda_k; \theta_k)\| \leq \sqrt{2\rho\alpha_k}, \end{aligned} \quad (3.3)$$

where the last inequality follows from Lemma 2.1 (iii). Since $g_\rho(\cdot; \theta^*)$ is the Moreau regularization of $g_0(\cdot; \theta^*)$, we have that $\lambda^* \in \operatorname{argmax}_\lambda g_\rho(\lambda, \theta^*)$ for all $\rho > 0$. Hence, $\nabla_\lambda g_\rho(\lambda^*; \theta^*) = 0$ and $\lambda^* = \pi_\rho(\lambda^*, \theta^*)$. From this observation, we obtain the bound below:

$$\begin{aligned} \|\pi_\rho(\lambda_k, \theta_k) - \lambda^*\| &= \|\pi_\rho(\lambda_k, \theta_k) - \pi_\rho(\lambda^*, \theta^*)\| \\ &\leq \|\pi_\rho(\lambda_k, \theta_k) - \pi_\rho(\lambda_k, \theta^*)\| + \|\pi_\rho(\lambda_k, \theta^*) - \pi_\rho(\lambda^*, \theta^*)\| \\ &= \rho \|\nabla_\lambda g_\rho(\lambda_k, \theta_k) - \nabla_\lambda g_\rho(\lambda_k, \theta^*)\| + \|\pi_\rho(\lambda_k, \theta^*) - \pi_\rho(\lambda^*, \theta^*)\| \\ &\leq \rho\kappa \|\theta_k - \theta^*\| + \|\lambda_k - \lambda^*\|, \end{aligned} \quad (3.4)$$

which follows from the Lipschitz continuity of $\nabla_\lambda g_\rho(\lambda; \theta)$ in θ uniformly in λ , the nonexpansivity of π_ρ in λ (Lemma 3.1). Hence, from (3.3) and (3.4), we obtain for all $i \geq 0$ that

$$\|\lambda_{i+1} - \lambda^*\| \leq \sqrt{2\rho\alpha_i} + \rho\kappa \|\theta_i - \theta^*\| + \|\lambda_i - \lambda^*\|.$$

For $k \geq 1$, by summing the above inequality over $i = 0, \dots, k-1$, we get

$$\begin{aligned} \|\lambda_k - \lambda^*\| &\leq \sum_{i=0}^{k-1} \left(\sqrt{2\rho\alpha_i} + \rho\kappa \|\theta_i - \theta^*\| \right) + \|\lambda_0 - \lambda^*\| \\ &\leq \sqrt{2\rho} \sum_{i=0}^{\infty} \sqrt{\alpha_i} + \rho\kappa \frac{\|\theta_0 - \theta^*\|}{1-\tau} + \|\lambda_0 - \lambda^*\|. \end{aligned} \quad (3.5)$$

□

Remark 3.1. *It is worth emphasizing that the bound C_λ can be tightened when θ^* is known. Indeed, when $\theta_0 = \theta^*$, the second term disappears.*

Next, we prove that the augmented Lagrangian scheme generates a sequence $\{\lambda_k\}$ such that $\bar{\lambda}_k \rightarrow \lambda^*$ as $k \rightarrow \infty$ by deriving a rate statement on the averaged sequence.

THEOREM 3.3 (Bound on dual suboptimality). *Let Assumptions 1 – 3 hold and let $\{\lambda_k\}_{k \geq 1}$ denote the sequence generated by Algorithm 1. In addition, let $\bar{\lambda}_k \triangleq \frac{1}{k} \sum_{i=1}^k \lambda_i$. Then it follows that for all $k \geq 1$:*

$$0 \leq f^* - g_\rho(\bar{\lambda}_k; \theta^*) = \sup_{\lambda} g_\rho(\lambda; \theta^*) - g_\rho(\bar{\lambda}_k; \theta^*) \leq \frac{B_g}{k}, \quad (3.6)$$

where $\lambda^* \in \operatorname{argmax}_\lambda g_0(\lambda, \theta^*)$, C_λ is defined in Theorem 3.2, and B_g is defined as follows:

$$B_g \triangleq \frac{1}{2\rho} \|\lambda_0 - \lambda^*\|^2 + C_\lambda \left(\sqrt{\frac{2}{\rho}} \sum_{k=0}^{\infty} \sqrt{\alpha_k} + \frac{\kappa \|\theta_0 - \theta^*\|}{1-\tau} \right).$$

Proof. From Lemma 2.1 and by recalling that the duality gap for $\mathcal{C}(\theta^*)$ is zero, it follows that $f^* = \max_\lambda g_\rho(\lambda; \theta^*)$ for all $\rho > 0$. By invoking the Lipschitz continuity of $\nabla_\lambda g_\rho(\lambda, \theta^*)$ in λ with constant $1/\rho$, the following holds for $i \geq 0$:

$$-g_\rho(\lambda_{i+1}; \theta^*) \leq -g_\rho(\lambda_i; \theta^*) - \nabla_\lambda g_\rho(\lambda_i; \theta^*)^\top (\lambda_{i+1} - \lambda_i) + \frac{1}{2\rho} \|\lambda_{i+1} - \lambda_i\|^2. \quad (3.7)$$

Under the concavity of $g_\rho(\lambda; \theta^*)$ in λ , we have that

$$-g_\rho(\lambda^*; \theta^*) \geq -g_\rho(\lambda_i; \theta^*) - \nabla_\lambda g_\rho(\lambda_i; \theta^*)^\top (\lambda^* - \lambda_i).$$

By combining the above inequality and (3.7), we get

$$\begin{aligned}
-g_\rho(\lambda_{i+1}; \theta^*) &\leq -g_\rho(\lambda_i; \theta^*) - \nabla_\lambda g_\rho(\lambda_i; \theta^*)^\top (\lambda_{i+1} - \lambda_i) + \frac{1}{2\rho} \|\lambda_{i+1} - \lambda_i\|^2 \\
&\leq -g_\rho(\lambda^*; \theta^*) - \nabla_\lambda g_\rho(\lambda_i; \theta^*)^\top (\lambda_{i+1} - \lambda^*) + \frac{1}{2\rho} \|\lambda_{i+1} - \lambda_i\|^2 \\
&= -g_\rho(\lambda^*; \theta^*) - \nabla_\lambda \mathcal{L}_\rho(x_{i+1}, \lambda_i; \theta_i)^\top (\lambda_{i+1} - \lambda^*) \\
&\quad + \delta_i^\top (\lambda_{i+1} - \lambda^*) + s_i^\top (\lambda_{i+1} - \lambda^*) + \frac{1}{2\rho} \|\lambda_{i+1} - \lambda_i\|^2 \\
&\leq -g_\rho(\lambda^*; \theta^*) - \frac{1}{\rho} (\lambda_{i+1} - \lambda_i)^\top (\lambda_{i+1} - \lambda^*) \\
&\quad + \frac{1}{2\rho} \|\lambda_{i+1} - \lambda_i\|^2 + \|\delta_i\| \|\lambda_{i+1} - \lambda^*\| + \|s_i\| \|\lambda_{i+1} - \lambda^*\|, \tag{3.8}
\end{aligned}$$

where $\delta_i \triangleq \nabla_\lambda g_\rho(\lambda_i; \theta_i) - \nabla_\lambda g_\rho(\lambda_i; \theta^*)$ and $s_i \triangleq \nabla_\lambda \mathcal{L}_\rho(x_{i+1}, \lambda_i; \theta_i) - \nabla_\lambda g_\rho(\lambda_i; \theta_i)$. By noting that $\|\lambda_{i+1} - \lambda_i\|^2 + 2(\lambda_{i+1} - \lambda_i)^\top (\lambda^* - \lambda_{i+1}) = \|\lambda_i - \lambda^*\|^2 - \|\lambda_{i+1} - \lambda^*\|^2$, we can rewrite (3.8) as

$$\begin{aligned}
-g_\rho(\lambda_{i+1}; \theta^*) &\leq -g_\rho(\lambda^*; \theta^*) + (\|\delta_i\| + \|s_i\|) \|\lambda_{i+1} - \lambda^*\| \\
&\quad + \frac{1}{2\rho} (\|\lambda_i - \lambda^*\|^2 - \|\lambda_{i+1} - \lambda^*\|^2). \tag{3.9}
\end{aligned}$$

By summing (3.9) over $i = 0, \dots, k-1$, replacing $g_\rho(\lambda^*; \theta^*)$ by $f^* = \sup_\lambda g_\rho(\lambda, \theta^*)$, we obtain

$$\begin{aligned}
-\sum_{i=0}^{k-1} \left(g_\rho(\lambda_{i+1}; \theta^*) - f^* \right) + \frac{1}{2\rho} \|\lambda_k - \lambda^*\|^2 &\leq \frac{1}{2\rho} \|\lambda_0 - \lambda^*\|^2 \\
&\quad + \sum_{i=0}^{k-1} (\|\delta_i\| + \|s_i\|) \|\lambda_{i+1} - \lambda^*\|. \tag{3.10}
\end{aligned}$$

Under concavity of $g_\rho(\lambda; \theta^*)$ in λ , the following holds:

$$-\left(g_\rho(\bar{\lambda}_k; \theta^*) - f^* \right) \leq -\frac{1}{k} \sum_{i=0}^{k-1} \left(g_\rho(\lambda_{i+1}; \theta^*) - f^* \right).$$

By dividing both sides of (3.10) by k and dropping the positive term on the left hand side, we get

$$f^* - g_\rho(\bar{\lambda}_k; \theta^*) \leq \frac{1}{k} \left(\frac{1}{2\rho} \|\lambda^*\|^2 + \sum_{i=0}^{k-1} (\|\delta_i\| + \|s_i\|) \|\lambda_{i+1} - \lambda^*\| \right).$$

Lemma 2.1(iii) and Proposition 2.4 imply that $\|s_i\| \leq \sqrt{\frac{2\alpha_i}{\rho}}$ and $\|\delta_i\| \leq \kappa \|\theta_i - \theta^*\|$, respectively for $i \geq 0$. In addition, from Theorem 3.2, we have $\|\lambda_i - \lambda^*\| \leq C_\lambda$ for all $i \geq 1$. Then by the summability of $\sqrt{\alpha_i}$, we have that

$$\sum_{i=0}^{\infty} (\|\delta_i\| + \|s_i\|) \|\lambda_{i+1} - \lambda^*\| \leq C_\lambda \left(\kappa \sum_{i=0}^{\infty} \|\theta_i - \theta^*\| + \sqrt{\frac{2}{\rho}} \sum_{i=0}^{\infty} \sqrt{\alpha_i} \right). \tag{3.11}$$

Furthermore, substituting $\sum_{i=0}^{\infty} \|\theta_i - \theta^*\| = \|\theta_0 - \theta^*\|/(1 - \tau)$ into (3.11) gives the desired bound and completes the proof. \square

Next, we derive a bound on the primal *infeasibility*, where the primal iterate sequence is computed such that Step 1 in Algorithm 1 is satisfied. Prior to proving our main result, we provide some supporting technical lemmas.

LEMMA 3.4. Assume that $\phi(\lambda) : \mathbb{R}^m \rightarrow \mathbb{R}$ is a concave function whose supremum is finite and is attained at λ_ϕ^* . In addition, assume that $\nabla\phi$ is Lipschitz continuous with constant L_ϕ . Then, we have that for all $\lambda \in \mathbb{R}^m$

$$\|\nabla\phi(\lambda)\| \leq \sqrt{2L_\phi(\phi(\lambda_\phi^*) - \phi(\lambda))}.$$

Proof. This is an immediate result of Theorem 2.1.5 in [23]. \square Next, we derive a bound on $d_{\mathcal{K}}(y + y')$ for any $y, y' \in \mathbb{R}^m$.

LEMMA 3.5. Given a closed convex cone \mathcal{K} , $d_{\mathcal{K}}(y + y') \leq d_{\mathcal{K}}(y) + \|y'\|$ for all $y, y' \in \mathbb{R}^m$.

Proof. From the definition of $d_{\mathcal{K}}(u)$ and the triangle inequality, we obtain the following bound:

$$\begin{aligned} d_{\mathcal{K}}(y + y') &= \|y + y' - \Pi_{\mathcal{K}}(y + y')\| \\ &= \|y - \Pi_{\mathcal{K}}(y) + y' + y - y + \Pi_{\mathcal{K}}(y) - \Pi_{\mathcal{K}}(y' + y)\| \\ &\leq \|y - \Pi_{\mathcal{K}}(y)\| + \|y' + y - \Pi_{\mathcal{K}}(y' + y) - (y - \Pi_{\mathcal{K}}(y))\|. \end{aligned}$$

Define $\Pi_{\mathcal{K}}^c(x) \triangleq x - \Pi_{\mathcal{K}}(x)$, which is a nonexpansive operator (cf. [10, Chapter 1.]). Using this operator, we can rewrite the above inequality as follows:

$$d_{\mathcal{K}}(y + y') \leq \|y - \Pi_{\mathcal{K}}(y)\| + \|\Pi_{\mathcal{K}}^c(y + y') - \Pi_{\mathcal{K}}^c(y)\| \leq d_{\mathcal{K}}(y) + \|y'\|,$$

where the last inequality follows from nonexpansivity of $\Pi_{\mathcal{K}}^c$. \square

We now derive the bound on the primal infeasibility.

THEOREM 3.6 (**Bound on primal infeasibility**). Let Assumptions 1–3 hold and let $\{\lambda_k\}_{k \geq 0}$ and $\{x_k\}_{k \geq 0}$ denote the sequences generated by Algorithm 1. Furthermore, let $\bar{x}_k \triangleq \frac{1}{k} \sum_{i=1}^k x_i$. Then, for all $k \geq 1$, it follows that

$$d_{-\mathcal{K}}(h(\bar{x}_k, \theta^*)) \leq \mathcal{V}(k) \triangleq \frac{C_1}{\sqrt{k}} + \frac{C_2}{k}, \quad (3.12)$$

where $C_1 \triangleq \sqrt{\frac{2B_g}{\rho} + \left(\frac{C_\lambda}{\rho}\right)^2}$ and $C_2 \triangleq \sqrt{\frac{2}{\rho}} \sum_{i=0}^{\infty} \sqrt{\alpha_i} + \frac{(L_h + \kappa)\|\theta_0 - \theta^*\|}{1 - \tau}$.

Proof. Let $u_i := \nabla_{\lambda} \mathcal{L}_\rho(x_{i+1}, \lambda_i; \theta_i)$ for all $i \geq 0$. Note that trivially

$$h(x_{i+1}; \theta^*) = h(x_{i+1}; \theta^*) + \underbrace{u_i - h(x_{i+1}; \theta_i) + \Pi_{-\mathcal{K}}\left(\frac{\lambda_i}{\rho} + h(x_{i+1}; \theta_i)\right)}_{\text{Term 1}},$$

since (2.6) implies that Term 1 = 0. Hence, using Lemma 3.5, we get the following inequality for all $i \geq 0$:

$$\begin{aligned} d_{-\mathcal{K}}(h(x_{i+1}; \theta^*)) &\leq d_{-\mathcal{K}}\left(\Pi_{-\mathcal{K}}\left(\frac{\lambda_i}{\rho} + h(x_{i+1}; \theta_i)\right)\right) \\ &\quad + \|u_i + h(x_{i+1}; \theta^*) - h(x_{i+1}; \theta_i)\|. \end{aligned} \quad (3.13)$$

Under Assumption 1(ii) (from Remark 2.1), we have $\|h(x_{i+1}, \theta_i) - h(x_{i+1}, \theta^*)\| \leq L_h \|\theta_i - \theta^*\|$ for all $i \geq 0$. Moreover, since $\Pi_{-\mathcal{K}}(y) \in -\mathcal{K}$ for all $y \in \mathbb{R}^m$ and $d_{-\mathcal{K}}(y) = 0$ for all $y \in -\mathcal{K}$, it follows from (3.13) that for all $i \geq 0$

$$d_{-\mathcal{K}}(h(x_{i+1}; \theta^*)) \leq \|u_i\| + L_{h,\theta} \|\theta^* - \theta_i\|.$$

Since $h(\cdot; \theta^*)$ is an affine function in x and $d_{-\mathcal{K}}(\cdot)$ is a convex function, their composition $d_{-\mathcal{K}}(h(\cdot; \theta^*))$ is a convex function as well; therefore, from Jensen's inequality we get

$$d_{-\mathcal{K}}\left(h(\bar{x}_k, \theta^*)\right) \leq \frac{1}{k} \sum_{i=0}^{k-1} d_{-\mathcal{K}}(h(x_{i+1}; \theta^*)) \leq \frac{1}{k} \sum_{i=0}^{k-1} (\|u_i\| + L_h \|\theta_i - \theta^*\|). \quad (3.14)$$

Recall that from Lemma 2.1 (iii), for $i = 0, \dots, k-1$,

$$\left\| \nabla_{\lambda} \mathcal{L}_{\rho}(x_{i+1}, \lambda_i; \theta_i) - \nabla_{\lambda} g_{\rho}(\lambda_i; \theta_i) \right\| \leq \sqrt{\frac{2\alpha_i}{\rho}};$$

therefore, we obtain that $\|u_i\| = \|\nabla_{\lambda} \mathcal{L}_{\rho}(x_{i+1}, \lambda_i; \theta_i)\| \leq \|\nabla_{\lambda} g_{\rho}(\lambda_i; \theta_i)\| + \sqrt{2\alpha_i/\rho}$. In addition, since $\|\nabla_{\lambda} g_{\rho}(\lambda_i; \theta_i)\| \leq \|\nabla_{\lambda} g_{\rho}(\lambda_i; \theta^*)\| + \kappa \|\theta_i - \theta^*\|$, we get the following bound:

$$\|u_i\| \leq \|\nabla_{\lambda} g_{\rho}(\lambda_i; \theta^*)\| + \sqrt{2\alpha_i/\rho} + \kappa \|\theta_i - \theta^*\|.$$

On the other hand, by Lemma 3.4, we have

$$\|\nabla_{\lambda} g_{\rho}(\lambda_i; \theta^*)\| \leq \sqrt{\frac{2}{\rho} (f^* - g_{\rho}(\lambda_i; \theta^*))}.$$

Combining this with the previous inequality leads to

$$\|u_i\| \leq \sqrt{\frac{2}{\rho} (f^* - g_{\rho}(\lambda_i; \theta^*))} + \sqrt{\frac{2\alpha_i}{\rho}} + \kappa \|\theta_i - \theta^*\|.$$

By substituting this bound into (3.14), we get that

$$\begin{aligned} d_{-\mathcal{K}}\left(h(\bar{x}_k, \theta^*)\right) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \sqrt{\frac{2}{\rho} (f^* - g_{\rho}(\lambda_i; \theta^*))} + \frac{1}{k} \left(\sum_{i=0}^{k-1} \sqrt{\frac{2\alpha_i}{\rho}} + (L_h + \kappa) \sum_{i=0}^{k-1} \|\theta_i - \theta^*\| \right) \\ &\leq \sqrt{\frac{2}{\rho} \left(f^* - \frac{1}{k} \sum_{i=0}^{k-1} g_{\rho}(\lambda_i; \theta^*) \right)} + \frac{1}{k} \left(\sum_{i=0}^{k-1} \sqrt{\frac{2\alpha_i}{\rho}} + (L_h + \kappa) \sum_{i=0}^{k-1} \|\theta_i - \theta^*\| \right), \end{aligned} \quad (3.15)$$

where the last inequality follows from concavity of square-root function $\sqrt{\cdot}$. The first term in (3.15) can be bounded using (3.10) and (3.11), which states that

$$f^* - \frac{1}{k} \sum_{i=0}^{k-1} g_{\rho}(\lambda_i; \theta^*) \leq \frac{1}{k} (B_g + g_{\rho}(\lambda_0; \theta^*) - g_{\rho}(\lambda_k; \theta^*)). \quad (3.16)$$

Note that $g_{\rho}(\lambda_0; \theta^*) - f^* \leq 0$, and using Lipschitz continuity of ∇g_{ρ} , we have $f^* - g_{\rho}(\lambda_k; \theta^*) \leq \frac{1}{2\rho} \|\lambda_k - \lambda^*\|^2 \leq \frac{1}{2\rho} C_{\lambda}^2$. The remaining terms in (3.15) can also be bounded as follows

$$\frac{1}{k} \left(\sum_{i=0}^{k-1} \sqrt{\frac{2\alpha_i}{\rho}} + (L_h + \kappa) \sum_{i=0}^{k-1} \|\theta_i - \theta^*\| \right) \leq \frac{1}{k} \left[\sum_{i=0}^{\infty} \sqrt{\frac{2\alpha_i}{\rho}} + \frac{(L_h + \kappa) \|\theta_0 - \theta^*\|}{1 - \tau} \right]. \quad (3.17)$$

The result follows by incorporating these bounds into (3.15). \square

We now proceed to derive lower and upper bounds on $f(\bar{x}_k, \theta^*) - f^*$. In contrast with standard unconstrained convex optimization, $f(\bar{x}_k, \theta^*)$ could be less than f^* , due to infeasibility of \bar{x}_k .

THEOREM 3.7 (Bounds on primal suboptimality). *Let Assumption 1–3 hold and let $\{x_k\}$ and $\{\lambda_k\}$ be the sequences generated by Algorithm 1. In addition, let $\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_k$. Then for all $k \geq 1$, the following hold:*

$$f(\bar{x}_k; \theta^*) - f^* \geq -\frac{\rho}{2} \mathcal{V}^2(k) - \|\lambda^*\| \mathcal{V}(k), \quad (3.18)$$

$$f(\bar{x}_k; \theta^*) - f^* \leq \frac{U}{k}, \quad (3.19)$$

for any $\lambda^* \in \operatorname{argmax} g_0(\lambda, \theta^*)$, where $\mathcal{V}(k)$ is defined in Theorem 3.6, $U \triangleq \sum_{i=0}^{\infty} \alpha_i + \frac{1}{2} \|\lambda_0\|^2 + \frac{\rho}{2} L_h^2 \frac{\|\theta_0 - \theta^*\|^2}{1 - \tau^2} + (\bar{C} L_h + 2L_f) \frac{\|\theta_0 - \theta^*\|}{1 - \tau}$, and $\bar{C} \triangleq C_\lambda + \|\lambda^*\|$.

Proof. We first prove the lower bound and then the upper bound.

Proof of the lower bound: Since $\sup_\lambda g_\rho(\lambda; \theta^*) = \min_{x \in X} \mathcal{L}_\rho(x, \lambda^*; \theta^*) = f^*$, we have that for all $k \geq 1$,

$$\begin{aligned} f^* &\leq \mathcal{L}_\rho(\bar{x}_k, \lambda^*; \theta^*) \\ &= f(\bar{x}_k; \theta^*) + \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(\bar{x}_k; \theta^*) + \frac{\lambda^*}{\rho} \right) - \frac{\|\lambda^*\|^2}{2\rho} \\ &\leq f(\bar{x}_k; \theta^*) + \frac{\rho}{2} \left(d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) + \frac{\|\lambda^*\|}{\rho} \right)^2 - \frac{\|\lambda^*\|^2}{2\rho}, \end{aligned}$$

where the first equality is a consequence of (2.5) while the second inequality follows from Lemma 3.5. By expanding the second term above inequality, we obtain

$$\begin{aligned} f^* &\leq f(\bar{x}_k; \theta^*) + \frac{\rho}{2} d_{-\mathcal{K}}^2(h(\bar{x}_k; \theta^*)) + d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \|\lambda^*\| \\ &\leq f(\bar{x}_k; \theta^*) + \frac{\rho}{2} \mathcal{V}^2(k) + \|\lambda^*\| \mathcal{V}(k), \end{aligned}$$

where the last inequality follows from Theorem 3.6.

Proof of the upper bound: Let x^* be an optimal solution to $\mathcal{C}(\theta^*)$, i.e., $x^* \in \mathcal{X}^*(\theta^*)$. Step 1 of Algorithm 1 implies that for all $i \geq 0$

$$\mathcal{L}_\rho(x_{i+1}, \lambda_i; \theta_i) \leq \mathcal{L}_\rho(x^*, \lambda_i; \theta_i) + \alpha_i.$$

Hence, by the definition of \mathcal{L}_ρ , it follows that

$$\begin{aligned} &f(x_{i+1}; \theta_i) + \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(x_{i+1}; \theta_i) + \frac{\lambda_i}{\rho} \right) - \frac{\|\lambda_i\|^2}{2\rho} \\ &\leq f(x^*; \theta_i) + \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(x^*; \theta_i) + \frac{\lambda_i}{\rho} \right) - \frac{\|\lambda_i\|^2}{2\rho} + \alpha_i, \end{aligned}$$

which leads to

$$\begin{aligned} f(x_{i+1}; \theta_i) - f(x^*; \theta_i) &\leq \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(x^*; \theta_i) + \frac{\lambda_i}{\rho} \right) \\ &\quad - \frac{\rho}{2} d_{-\mathcal{K}}^2 \left(h(x_{i+1}; \theta_i) + \frac{\lambda_i}{\rho} \right) + \alpha_i. \end{aligned} \quad (3.20)$$

Step 2 of Algorithm 1 and (3.1) imply that

$$\begin{aligned}\lambda_{i+1} &= \rho \left(h(x_{i+1}; \theta_i) + \frac{\lambda_i}{\rho} - \Pi_{-\mathcal{K}} \left(\frac{\lambda_i}{\rho} + h(x_{i+1}; \theta_i) \right) \right) \\ \implies d_{-\mathcal{K}} \left(h(x_{i+1}; \theta_i) + \frac{\lambda_i}{\rho} \right) &= \frac{\|\lambda_{i+1}\|}{\rho}.\end{aligned}\quad (3.21)$$

In addition, by using Lemma 3.5, it follows that

$$d_{-\mathcal{K}} \left(h(x^*; \theta_i) + \frac{\lambda_i}{\rho} \right) \leq d_{-\mathcal{K}}(h(x^*; \theta_i)) + \frac{\|\lambda_i\|}{\rho}. \quad (3.22)$$

Substituting (3.21) and (3.22) in (3.20), we obtain for all $i \geq 0$

$$\begin{aligned}f(x_{i+1}; \theta_i) - f(x^*; \theta_i) &\leq \frac{\rho}{2} \left(d_{-\mathcal{K}}(h(x^*; \theta_i)) + \frac{\|\lambda_i\|}{\rho} \right)^2 - \frac{1}{2\rho} \|\lambda_{i+1}\|^2 + \alpha_i \\ &= \frac{\rho}{2} d_{-\mathcal{K}}^2(h(x^*; \theta_i)) + \|\lambda_i\| d_{-\mathcal{K}}(h(x^*; \theta_i)) \\ &\quad + \frac{1}{2\rho} (\|\lambda_i\|^2 - \|\lambda_{i+1}\|^2) + \alpha_i.\end{aligned}\quad (3.23)$$

According to Remark 2.1, we have $\|h(x^*; \theta_i) - h(x^*; \theta^*)\| \leq L_h \|\theta_i - \theta^*\|$; hence, we have $h(x^*; \theta_i) = h(x^*; \theta^*) + v_i$ for some $v_i \in L_{h,\theta} \|\theta_i - \theta^*\| \mathcal{B}(\mathbf{0}, 1)$ for all $i \geq 0$. Using Lemma 3.5, for each $i \geq 0$, we get $d_{-\mathcal{K}}(h(x^*; \theta_i)) \leq d_{-\mathcal{K}}(h(x^*; \theta^*)) + \|v_i\|$; thus, since $x^* \in \mathcal{X}^*(\theta^*)$, we have $h(x^*; \theta^*) \preceq_{\mathcal{K}} \mathcal{K}$, and this implies

$$d_{-\mathcal{K}}(h(x^*; \theta_i)) \leq L_{h,\theta} \|\theta_i - \theta^*\|, \quad (3.24)$$

for $i \geq 0$. By substituting (3.24) into (3.23), we get for all $i \geq 0$

$$f(x_{i+1}; \theta_i) - f(x^*; \theta_i) \leq \frac{\rho}{2} L_h^2 \|\theta_i - \theta^*\|^2 + L_h \|\theta_i - \theta^*\| \|\lambda_i\| \quad (3.25)$$

$$\begin{aligned}&+ \frac{1}{2\rho} (\|\lambda_i\|^2 - \|\lambda_{i+1}\|^2) + \alpha_i \\ &\leq \frac{\rho}{2} L_h^2 \|\theta_i - \theta^*\|^2 + \bar{C} L_h \|\theta_i - \theta^*\| \\ &+ \frac{1}{2\rho} (\|\lambda_i\|^2 - \|\lambda_{i+1}\|^2) + \alpha_i,\end{aligned}\quad (3.26)$$

where the last inequality follows from $\|\lambda_i - \lambda^*\| \leq C_\lambda$ (Proposition 3.2), i.e., $\|\lambda_i\| \leq \bar{C} \triangleq C_\lambda + \|\lambda^*\|$ for all $i \geq 0$. Next, from the Lipschitz continuity of f in θ by Assumption 1(i), it follows that

$$f(x_{i+1}; \theta_i) - f(x^*; \theta_i) \geq f(x_{i+1}; \theta^*) - f(x^*; \theta^*) - 2L_f \|\theta_i - \theta^*\|.$$

Combining two above inequalities results in the following:

$$\begin{aligned}f(x_{i+1}; \theta^*) - f^* &\leq \frac{\rho}{2} L_h^2 \|\theta_i - \theta^*\|^2 + (\bar{C} L_h + 2L_f) \|\theta_i - \theta^*\| \\ &+ \frac{1}{2\rho} (\|\lambda_i\|^2 - \|\lambda_{i+1}\|^2) + \alpha_i.\end{aligned}$$

Summing the above inequality for $i = 0$ to $k-1$, we obtain the following:

$$\begin{aligned}\sum_{i=0}^{k-1} \left(f(x_{i+1}; \theta^*) - f^* \right) &\leq \sum_{i=0}^{\infty} \alpha_i + \frac{1}{2\rho} \|\lambda_0\|^2 + \frac{\rho}{2} L_h^2 \frac{\|\theta_0 - \theta^*\|^2}{1 - \tau^2} \\ &+ (\bar{C} L_h + 2L_f) \frac{\|\theta_0 - \theta^*\|}{1 - \tau}.\end{aligned}$$

Since $f(x; \theta^*)$ is convex in x , dividing both sides of the above inequality by k gives the desired result. \square

3.2. Convergence analysis for increasing $\{\rho_k\}$. In 1976, Rockafellar [28] proposed several different variants of inexact augmented Lagrangian schemes where the penalty parameter could be updated between iterations and under suitable summability conditions on the sequence $\{\alpha_k \rho_k\}$, it was established that the sequence of dual iterates $\{\lambda_k\}$ is bounded. In addition, upper bounds on primal suboptimality and infeasibility were derived. Recently, Aybat and Iyengar [4] extended this result to conic convex programs, provided both upper and lower bounds on the suboptimality, and presented sequences $\{\rho_k\}$ and $\{\alpha_k\}$ under which the primal function converges linearly to its optimal value. Necoara et al. [21] also considered similar inexact schemes for conic convex programs where the penalty parameter ρ_k is tuned adaptively. In this scheme, they apply a search procedure which finds an upper bound on λ^* in logarithmic number of steps, by performing a single outer iteration and restarting the augmented Lagrangian method. In what follows, we analyze the convergence properties of Algorithm 1 for an increasing penalty parameter sequence.

The proposed sequences of ρ_k and α_k in Theorem 3.11 are extensions of those presented in [4] and take learning into consideration. We initiate the analysis on the rate of convergence by first deriving the bound on primal infeasibility, which is subsequently used later to derive bounds on primal suboptimality. These statements are then utilized in deriving the overall computational complexity in Section 4.2. Suppose $\{\rho_k\}_{k \geq 0} \subset \mathbb{R}_{++}$, and recall that according to (2.5),

$$\mathcal{L}_{\rho_k}(x, \lambda; \theta) = f(x; \theta) + \frac{\rho_k}{2} d_{\mathcal{K}}^2 \left(h(x; \theta) + \frac{\lambda}{\rho_k} \right) - \frac{\|\lambda\|^2}{2\rho_k}, \quad (3.27)$$

and i.e., $g_{\rho_k}(\lambda; \theta) = \inf_{x \in X} \mathcal{L}_{\rho_k}(x, \lambda; \theta)$. Since $g_{\rho}(\cdot; \theta)$ is the Moreau envelope of $g_0(\cdot; \theta)$ for any $\rho > 0$, both $g_{\rho_k}(\cdot; \theta)$ and $g_0(\cdot; \theta)$ have the same set of maximizers for any fixed $\theta \in \Theta$, i.e., $\operatorname{argmax}_{\lambda} g_{\rho_k}(\lambda; \theta) = \operatorname{argmax}_{\lambda} g_0(\lambda; \theta)$, for all $k \geq 0$. Moreover, it is also true that $\max_{\lambda} g_{\rho_k}(\lambda; \theta) = \max_{\lambda} g_0(\lambda; \theta)$; hence, $f^* = \max_{\lambda} g_{\rho_k}(\lambda; \theta^*)$ for all $k \geq 0$. We begin this section with a lemma, similar to Proposition 3.2, which proves that the dual iterate sequence $\{\lambda_k\}_{k \in \mathbb{Z}_+}$ stays bounded.

LEMMA 3.8. *Let Assumptions 1 and 2 hold. Given $\lambda_0 \in \mathcal{K}^*$, let $\{\lambda_k\}_{k \in \mathbb{Z}_+}$ be the dual iterate sequence generated by Algorithm 1 when the following conditions hold:*

- i) $\rho_k = \rho_0 \beta^k$, where ρ_0 and β are positive scalars such that $\beta > 1$,
- ii) $\beta\tau < 1$, where $\tau \in (0, 1)$ is the constant defined in Assumption 2,
- iii) $\sum_{k=0}^{\infty} (\alpha_k \rho_k)^{\frac{1}{2}} < \infty$.

Then, for all $k \geq 0$, it follows that $\|\lambda_k - \lambda^\| \leq C'_{\lambda}$, where C'_{λ} is defined as*

$$C'_{\lambda} \triangleq \sum_{i=0}^{\infty} \sqrt{2\alpha_i \rho_i} + \frac{\rho_0 \kappa \|\theta_0 - \theta^*\|}{1 - \beta\tau} + \|\lambda_0 - \lambda^*\|,$$

where λ^ is any point in $\Lambda^* \triangleq \operatorname{argmax}_{\lambda} g_0(\lambda, \theta^*)$.*

Proof. We follow the same lines of proof in Proposition 3.2 for an arbitrary given penalty sequence $\{\rho_k\} \subset \mathbb{R}_{++}$ to obtain a slightly modified version of (3.5) as follows

$$\|\lambda_k - \lambda^*\| \leq \sum_{i=0}^{k-1} \sqrt{2\rho_i \alpha_i} + \kappa \sum_{i=0}^{k-1} \rho_i \|\theta_i - \theta^*\| + \|\lambda_0 - \lambda^*\|. \quad (3.28)$$

For $i \geq 0$, since $\rho_i = \rho_0 \beta^i$ by hypothesis, and $\|\theta_i - \theta^*\| \leq \tau^i \|\theta_0 - \theta^*\|$ by Assumption 2,

we can upper bound (3.28) as follows using the condition $\beta\tau < 1$:

$$\begin{aligned}\|\lambda_{k+1} - \lambda^*\| &\leq \sum_{i=0}^{\infty} \sqrt{2\rho_i\alpha_i} + \rho_0\kappa\|\theta_0 - \theta^*\| \sum_{i=0}^{\infty} (\beta\tau)^i + \|\lambda_0 - \lambda^*\|, \\ &= \sum_{i=0}^{\infty} \sqrt{2\rho_i\alpha_i} + \frac{\rho_0\kappa\|\theta_0 - \theta^*\|}{1 - \beta\tau} + \|\lambda_0 - \lambda^*\|.\end{aligned}$$

□

Next, we derive a bound on the primal infeasibility.

PROPOSITION 3.9. *Under Assumption 1, let $\{x_k, \lambda_k\}_{k \in \mathbb{Z}_+}$ be the primal-dual iterate sequence generated by Algorithm 1 for a given penalty sequence $\{\rho_k\}_{k \in \mathbb{Z}}$. Then, for all $k \geq 0$, it follows that*

$$d_{-\mathcal{K}}(h(x_{k+1}; \theta^*)) \leq \frac{\|\lambda_{k+1} - \lambda_k\|}{\rho_k} + L_h\|\theta_k - \theta^*\|.$$

Proof. According to Step 2 of Algorithm 1, in particular from (3.1), we have that

$$h(x_{k+1}; \theta_k) = \frac{\lambda_{k+1} - \lambda_k}{\rho_k} + \Pi_{-\mathcal{K}}\left(h(x_{k+1}; \theta_k) + \frac{\lambda_k}{\rho_k}\right). \quad (3.29)$$

Therefore, Lemma 3.5 implies that

$$\begin{aligned}d_{-\mathcal{K}}(h(x_{k+1}; \theta_k)) &\leq \frac{\|\lambda_{k+1} - \lambda_k\|}{\rho_k} + d_{-\mathcal{K}}\left(\Pi_{-\mathcal{K}}\left(h(x_{k+1}; \theta_k) + \frac{\lambda_k}{\rho_k}\right)\right) \\ &\leq \frac{\|\lambda_{k+1} - \lambda_k\|}{\rho_k},\end{aligned} \quad (3.30)$$

where (3.30) follows from $d_{-\mathcal{K}}\left(\Pi_{-\mathcal{K}}\left(h(x_k; \theta_k) + \frac{\lambda_k}{\rho_k}\right)\right) = 0$ since $\Pi_{-\mathcal{K}}(y) \in -\mathcal{K}$ for all $y \in \mathbb{R}^m$ and $d_{-\mathcal{K}}(y) = 0$ for all $y \in -\mathcal{K}$. Under Assumption 1(ii), in particular from Remark 2.1, we also have $\|h(x_{i+1}, \theta_i) - h(x_{i+1}, \theta^*)\| \leq L_h\|\theta_i - \theta^*\|$ for all $i \geq 0$. Therefore, from the triangular inequality of Lemma 3.5, it follows that

$$d_{-\mathcal{K}}(h(x_{k+1}; \theta^*)) \leq d_{-\mathcal{K}}(h(x_{k+1}; \theta_k)) + L_h\|\theta_k - \theta^*\| \leq \frac{\|\lambda_{k+1} - \lambda_k\|}{\rho_k} + L_h\|\theta_k - \theta^*\|.$$

□

Our next result provides bounds on the primal sub-optimality.

PROPOSITION 3.10. *Under Assumption 1, let $\{x_k, \lambda_k\}_{k \in \mathbb{Z}_+}$ be the primal-dual iterate sequence generated by Algorithm 1 for a given penalty sequence $\{\rho_k\}_{k \in \mathbb{Z}}$. Then, for all $k \geq 1$, it follows that*

$$f(x_{k+1}; \theta^*) - f^* \geq -\frac{1}{\rho_k} (\|\lambda_{k+1}\| + \|\lambda_k - \lambda^*\|)^2 - \rho_k L_h^2 \|\theta_k - \theta^*\|^2, \quad (3.31)$$

$$f(x_{k+1}; \theta^*) - f^* \leq \frac{\|\lambda_k\|^2}{\rho_k} + 2L_f\|\theta_k - \theta^*\| + \rho_k L_h^2 \|\theta_k - \theta^*\|^2 + \alpha_k. \quad (3.32)$$

where λ^* is any point in $\Lambda^* \triangleq \operatorname{argmax}_{\lambda} g_0(\lambda; \theta^*)$.

Proof. We start by proving the lower bound in (3.31), and then prove the upper bound in (3.32).

Proof of (3.31): Since $f^* = \max_{\lambda} g_{\rho_k}(\lambda; \theta^*)$, $\lambda \in \Lambda^*$ implies $\lambda^* \in \arg\max g_{\rho_k}(\lambda; \theta^*)$. Consequently, we have $f^* = \min_{x \in X} \mathcal{L}_{\rho_k}(x, \lambda^*; \theta^*)$. Hence, for all $k \geq 0$,

$$f^* \leq \mathcal{L}_{\rho_k}(x_{k+1}, \lambda^*; \theta^*) = f(x_{k+1}; \theta^*) + \frac{\rho_k}{2} d_{-\mathcal{K}}^2 \left(h(x_{k+1}; \theta^*) + \frac{\lambda^*}{\rho_k} \right) - \frac{\|\lambda^*\|^2}{2\rho_k}. \quad (3.33)$$

Once again the using the triangular inequality of Lemma 3.5, we get

$$\begin{aligned} d_{-\mathcal{K}} \left(h(x_{k+1}; \theta^*) + \frac{\lambda^*}{\rho_k} \right) &\leq d_{-\mathcal{K}} \left(h(x_{k+1}; \theta_k) + \frac{\lambda_k}{\rho_k} \right) \\ &\quad + \left\| h(x_{k+1}; \theta^*) - h(x_{k+1}; \theta_k) + \frac{\lambda^* - \lambda_k}{\rho_k} \right\| \\ &\leq \left\| \frac{\lambda_{k+1}}{\rho_k} \right\| + \left\| h(x_{k+1}; \theta_k) - h(x_{k+1}; \theta^*) + \frac{\lambda_k - \lambda^*}{\rho_k} \right\| \\ &= \frac{1}{\rho_k} \left(\|\lambda_{k+1}\| + \|\lambda_k - \lambda^*\| \right) + L_{h,\theta} \|\theta_k - \theta^*\|, \end{aligned} \quad (3.34)$$

where in the second inequality, we use the identity $\|\lambda_{k+1}\| = \rho_k d_{-\mathcal{K}}(h(x_{k+1}; \theta_k) + \frac{\lambda_k}{\rho_k})$, which follows from (3.30), and in the third inequality, we invoke the Lipschitz continuity of function $h(x; \theta)$ in θ (see Remark 2.1). Hence, using the identity $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ on (3.34) to bound $d_{-\mathcal{K}}^2 \left(h(x_{k+1}; \theta^*) + \frac{\lambda^*}{\rho_k} \right)$, and using the resulting bound within (3.33), we obtain the desired lower bound in (3.31).

Proof of (3.32): Step 1 of Algorithm 1 implies that

$$\mathcal{L}_{\rho_k}(x_{k+1}, \lambda_k; \theta_k) \leq \inf_{x \in X} \mathcal{L}_{\rho_k}(x, \lambda_k; \theta_k) + \alpha_k \leq \mathcal{L}_{\rho_k}(x^*, \lambda_k; \theta_k) + \alpha_k.$$

By the definition of \mathcal{L}_{ρ_k} in (3.27), and using the fact that $d_{-\mathcal{K}}(y) \geq 0$ for any $y \in \mathbb{R}^m$, we have that

$$f(x_{k+1}; \theta_k) - f(x^*; \theta_k) \leq \frac{\rho_k}{2} d_{-\mathcal{K}}^2 \left(h(x^*; \theta_k) + \frac{\lambda_k}{\rho_k} \right) + \alpha_k. \quad (3.35)$$

Note that using the triangular inequality of Lemma 3.5 twice, we get

$$\begin{aligned} d_{-\mathcal{K}} \left(h(x^*; \theta_k) + \frac{\lambda_k}{\rho_k} \right) &\leq d_{-\mathcal{K}} \left(h(x^*; \theta^*) + \frac{\lambda_k}{\rho_k} \right) + \|h(x^*; \theta_k) - h(x^*; \theta^*)\| \\ &\leq d_{-\mathcal{K}}(h(x^*; \theta^*)) + \frac{\|\lambda_k\|}{\rho_k} + \|h(x^*; \theta_k) - h(x^*; \theta^*)\| \\ &\leq \frac{\|\lambda_k\|}{\rho_k} + L_{h,\theta} \|\theta_k - \theta^*\|, \end{aligned} \quad (3.36)$$

where in the third inequality we used the fact that $d_{-\mathcal{K}}(h(x^*; \theta^*)) = 0$ along with Lipschitz continuity of function $h(x^*; \theta)$ in θ . By substituting (3.36) into (3.35) and again using the identity $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain the following:

$$f(x_{k+1}; \theta_k) - f(x^*; \theta_k) \leq \rho_k L_h^2 \|\theta_k - \theta^*\|^2 + \frac{\|\lambda_k\|^2}{\rho_k} + \alpha_k. \quad (3.37)$$

Since $f(x; \theta)$ is Lipschitz continuous in θ for all $x \in X$ by Assumption 1(i), after bounding each difference term in the right hand side of the following equality,

$$\begin{aligned} f(x_{k+1}; \theta^*) - f^* &= f(x_{k+1}; \theta^*) - f(x_{k+1}; \theta_k) + f(x_{k+1}; \theta_k) - f(x^*; \theta_k) \\ &\quad + f(x^*; \theta_k) - f(x^*; \theta^*). \end{aligned}$$

and using (3.37) we obtain the desired upper bound given in (3.32). \square

We conclude this subsection with a formal rate statement for the sub-optimality and the infeasibility of the sequence $\{x_k\}$.

THEOREM 3.11. *Under Assumption 1 and 2, let $\{x_k, \lambda_k\}_{k \in \mathbb{Z}_+}$ be the primal-dual iterate sequence generated by Algorithm 1 for the increasing penalty sequence $\{\rho_k\}_{k \in \mathbb{Z}}$ and inexact optimality parameter sequence $\{\alpha_k\}_{k \in \mathbb{Z}}$ defined as follows: given some $c, \alpha_0, \rho_0 > 0$, $\beta > 1$ and $\tau \in (0, 1)$ such that $\delta \triangleq \beta\tau < 1$, let*

$$\alpha_k \triangleq \frac{\alpha_0}{(k+1)^{2(1+c)}\beta^k}, \quad \rho_k \triangleq \rho_0\beta^k, \quad k \geq 0.$$

Then, for all $k \geq 0$, the following bounds hold:

$$|f(x_{k+1}, \theta^*) - f^*| \leq \frac{B_k}{\beta^k}, \quad (3.38)$$

$$d_{-\mathcal{K}}(h(x_{k+1}; \theta^*)) \leq \frac{1}{\beta^k} \left(\frac{2C'_\lambda}{\rho_0} + L_{h,\theta} \|\theta_0 - \theta^*\| \delta^k \right), \quad (3.39)$$

where C'_λ is the constant defined in Lemma 3.8 and B_k is defined as follows

$$B_k \triangleq \frac{1}{\rho_0} (2C'_\lambda + \|\lambda^*\|)^2 + \rho_0 \left(L_{h,\theta} \|\theta_0 - \theta^*\| \delta^k + \frac{L_f}{\rho_0 L_{h,\theta}} \right)^2 + \frac{\alpha_0}{(k+1)^{2(1+c)}}. \quad (3.40)$$

Proof. Recall that from Assumption 2, for some $\tau \in (0, 1)$, we have $\|\theta_k - \theta^*\| \leq \|\theta_0 - \theta^*\| \tau^k$ for all $k \geq 0$. By hypothesis, $\rho_k = \rho_0 \beta^k$ for $k \geq 0$, $\delta \triangleq \beta\tau < 1$, and $\sum_{k=0}^{\infty} \sqrt{\alpha_k \rho_k} = \sum_{k=0}^{\infty} \frac{\alpha_0}{k^{1+c}} < \infty$; therefore, the conditions of Lemma 3.8 are satisfied and as a consequence, we have that $\|\lambda_k - \lambda^*\| \leq C'_\lambda$ for all $k \geq 0$. Next, by combining the lower and upper bounds on primal suboptimality obtained in Proposition 3.10, we obtain that for all $k \geq 0$

$$\begin{aligned} & \left| f(x_{k+1}, \theta^*) - f^* \right| \\ & \leq \max \left\{ \frac{\|\lambda_k\|^2}{\rho_k} + 2L_f \|\theta_k - \theta^*\| + \alpha_k, \frac{1}{\rho_k} \left(\|\lambda_{k+1}\| + \|\lambda_k - \lambda^*\| \right)^2 \right\} \\ & \quad + \rho_k L_h^2 \|\theta_k - \theta^*\|^2, \\ & \leq \frac{1}{\beta^k} \max \left\{ \frac{1}{\rho_0} (C'_\lambda + \|\lambda^*\|)^2 + 2L_f \|\theta_0 - \theta^*\| \delta^k + \frac{\alpha_0}{(k+1)^{2(1+c)}}, \frac{1}{\rho_0} (2C'_\lambda + \|\lambda^*\|)^2 \right\} \\ & \quad + \frac{1}{\beta^k} \rho_0 L_{h,\theta}^2 \|\theta_0 - \theta^*\|^2 \delta^{2k}, \\ & \leq \frac{1}{\beta^k} \left[\frac{1}{\rho_0} (2C'_\lambda + \|\lambda^*\|)^2 + \rho_0 L_{h,\theta}^2 \|\theta_0 - \theta^*\|^2 \delta^{2k} + 2L_f \|\theta_0 - \theta^*\| \delta^k + \frac{\alpha_0}{(k+1)^{2(1+c)}} \right]. \end{aligned}$$

Hence, completing the square in the last inequality, we obtain the desired result. To prove the rate statement for infeasibility, we use Proposition 3.9 and Lemma 3.8:

$$\begin{aligned} d_{-\mathcal{K}}(h(x_k; \theta^*)) & \leq \frac{\|\lambda_{k+1} - \lambda_k\|}{\rho_k} + L_{h,\theta} \|\theta_k - \theta^*\| \\ & \leq \frac{1}{\beta^k} \left(\frac{2C'_\lambda}{\rho_0} + L_{h,\theta} \|\theta_0 - \theta^*\| \delta^k \right). \end{aligned} \quad (3.41)$$

\square

4. Overall iteration complexity analysis. Implementing the inexact augmented Lagrangian algorithm involves performing inner and outer loops, where each outer loop corresponds to one update of the Lagrange multiplier according to Step 2 in Algorithm 1, while the inner loops correspond to iterations of the scheme employed to compute x_{k+1} as in Step 1. Hence, to assess the overall computational complexity of our inexact augmented Lagrangian approach, it is essential to specify the algorithm used for inner optimization. We assume that

$$f(x; \theta) := q(x; \theta) + p(x; \theta),$$

where the functions p and q represent the *smooth* and *nonsmooth* parts of f , respectively. Then, following the formulation (2.5),

$$\mathcal{L}_\rho(x, \lambda; \theta) = q(x; \theta) + \nu_\rho(x, \lambda; \theta), \quad (4.1)$$

$$\text{where } \nu_\rho(x, \lambda; \theta) \triangleq p(x; \theta) + \frac{\rho}{2} d_{-\mathcal{K}}^2\left(\frac{\lambda}{\rho} + h(x; \theta)\right) - \frac{\|\lambda\|^2}{2\rho}.$$

In the representation (4.1), the function q captures the nonsmooth part of augmented Lagrangian function while the function ν_ρ represents the smooth part. This is a particular case of the *composite* convex minimization problem studied in [5, 22, 30]. In [5, 22], the authors developed Accelerated Proximal Gradient (APG) methods, inspired by Nesterov's optimal scheme [23], that can compute an ϵ -optimal solution to composite convex optimization problems within $\mathcal{O}(1/\sqrt{\epsilon})$ iterations.

In what follows, we assume that the inner loop is resolved by a particular implementation of the APG algorithm called Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [5]. Implementing this scheme requires that $\nabla_x \nu_\rho(x; \theta)$ be Lipschitz continuous in x uniformly in θ . In this respect, the next lemma states the conditions under which we may have such a property. Before presenting this result, we make the following assumptions.

Assumption 4. Let $q, p : X \times \Theta \rightarrow \mathbb{R}$ be proper, closed, convex functions of x for all $\theta \in \Theta$ such that $p(x; \theta)$ is differentiable in x on an open set containing X for any fixed $\theta \in \Theta$. Moreover, $\nabla_x p(x; \theta)$ is Lipschitz continuous in x , uniformly for all $\theta \in \Theta$ with constant $L_{p,x}$. Recall that under Assumption 1(ii), function $h(x; \theta)$ is Lipschitz continuous in x , uniformly for all $\theta \in \Theta$, with constant $L_{h,x} \triangleq \max_{\theta \in \Theta} \|A(\theta)\|$.

LEMMA 4.1 (Lipschitz continuity of $\nabla_x \nu_\rho(x; \theta)$). Under Assumption 4, for any given $\theta \in \Theta$, the gradient function $\nabla_x \nu_\rho(x; \theta)$ is Lipschitz continuous in x with constant $L_{\nu,x}(\rho, \theta)$ defined as

$$L_{\nu,x}(\rho, \theta) \triangleq L_{p,x} + \rho \|A(\theta)\|^2 \leq L_{p,x} + \rho L_{h,x}^2.$$

Proof. Recall that $h(x; \theta) = A(\theta)x + b(\theta)$ is an affine function in x for all $\theta \in \Theta$. Since for any closed convex cone $\mathcal{K} \in \mathbb{R}^m$, $\Pi_{\mathcal{K}^*}(x) = x - \Pi_{-\mathcal{K}}(x)$ for all $x \in \mathbb{R}^m$ and $\nabla d_{-\mathcal{K}}^2(y) = 2(y - \Pi_{-\mathcal{K}}(y))$ for all $y \in \mathbb{R}^m$, the following holds:

$$\begin{aligned} \nabla_x \nu(x, \lambda; \theta) &= \nabla_x p(x; \theta) + \rho A(\theta)^\top \left(h(x; \theta) + \frac{\lambda}{\rho} - \Pi_{-\mathcal{K}} \left(h(x; \theta) + \frac{\lambda}{\rho} \right) \right) \\ &= \nabla_x p(x; \theta) + \rho A(\theta)^\top \Pi_{\mathcal{K}^*} \left(h(x; \theta) + \frac{\lambda}{\rho} \right). \end{aligned}$$

Then, by adding and subtracting terms and by invoking the triangle inequality and Lipschitz continuity, it follows that for all $x, x' \in X$ and $\theta \in \Theta$,

$$\begin{aligned}
& \|\nabla_x \nu(x, \lambda; \theta) - \nabla_x \nu(x', \lambda; \theta)\| \\
&= \left\| \nabla_x p(x; \theta) + \rho A(\theta)^\top \Pi_{\mathcal{K}^*} \left(h(x; \theta) + \frac{\lambda}{\rho} \right) \right. \\
&\quad \left. - \nabla_x p(x'; \theta) - \rho A(\theta)^\top \Pi_{\mathcal{K}^*} \left(h(x'; \theta) + \frac{\lambda}{\rho} \right) \right\| \\
&\leq \|\nabla_x p(x; \theta) - \nabla_x p(x'; \theta)\| \\
&\quad + \rho \|A(\theta)\| \left\| \Pi_{\mathcal{K}^*} \left(h(x; \theta) + \frac{\lambda}{\rho} \right) - \Pi_{\mathcal{K}^*} \left(h(x'; \theta) + \frac{\lambda}{\rho} \right) \right\| \\
&\leq L_{p,x} \|x - x'\| + \rho \|A(\theta)\|^2 \|x - x'\|,
\end{aligned} \tag{4.2}$$

where the last inequality uses the nonexpansivity property of the projection operator. \square

Given x_k computed in the previous outer iteration, we use these inner iterations within Algorithm 1 to compute an inexact solution, x_{k+1} , with accuracy α_k to the following optimization problem

$$g_{\rho_k}(\lambda_k, \theta_k) = \min_{x \in X} q(x; \theta_k) + \nu_{\rho_k}(x, \lambda_k; \theta_k) \tag{4.4}$$

where x_{k+1} is defined as

$$q(x_{k+1}; \theta_k) + \nu_{\rho_k}(x_{k+1}, \lambda_k; \theta_k) \leq g_{\rho_k}(\lambda_k, \theta_k) + \alpha_k.$$

Algorithm 2 APG($x_k, \theta_k, \alpha_k, \rho_k$) (implemented on subproblem (4.4))

Set $z_0 \leftarrow x_k$, $y_1 \leftarrow z_0$, $m_1 \leftarrow 1$ and $t \leftarrow 1$. Then, for all $t \geq 1$, update:

1. $z_t \leftarrow \operatorname{argmin}_{z \in X} \left\{ q(z; \theta_k) + \nabla_x \nu_{\rho_k}(y_t, \lambda_k; \theta_k)^\top (z - y_t) + \frac{L_{\nu,x}(\rho_k, \theta_k)}{2} \|z - y_t\|^2 \right\}$
 2. $m_{t+1} \leftarrow (1 + \sqrt{1 + 4m_t^2})/2$
 3. $y_{t+1} \leftarrow z_t + \left(\frac{m_t - 1}{m_{t+1}} \right) (z_t - z_{t-1})$
 4. If $t \geq T_k \triangleq \sqrt{\frac{2L_{\nu,x}(\rho_k, \theta_k)}{\alpha_k}} D_x$, then STOP; else $t \leftarrow t + 1$ and go to Step 1.
-

To obtain an accuracy of α_k , at most $T_k \triangleq \sqrt{\frac{2L_{\nu,x}(\rho_k, \theta_k)}{\alpha_k}} D_x$ steps of the APG scheme are required at epoch k ; this follows from the iteration complexity result of APG (See in [5, Theorem 4.4] for a proof).

LEMMA 4.2. *Let q and ν functions satisfying Assumption 4. Fix $\alpha_k > 0$ and let $\{z_t, y_t\}$ denote the sequence of iterates generated by the APG algorithm. Then, $q(z_t, \theta_k) + \nu_{\rho_k}(z_t, \lambda_k; \theta_k) \leq g_{\rho_k}(\lambda_k, \theta_k) + \alpha_k$ whenever*

$$t \geq \sqrt{\frac{2L_{\nu,x}(\rho_k, \theta_k)}{\alpha_k}} \|x_k - x_{k+1}^*\| - 1,$$

where $x_{k+1}^* \in \operatorname{argmin}_{x \in X} q(x; \theta_k) + \nu_{\rho_k}(x, \lambda_k; \theta_k)$.

Algorithm 2 displays an APG algorithm studied in [5].

4.1. Overall iteration complexity for constant penalty ρ . Next, we derive the overall iteration complexity of Algorithm 1 in which APG is used for solving the subproblem in (4.4) to compute x_k satisfying Step 1 in Algorithm 1.

THEOREM 4.3. *Let Assumptions 1, 2, 3 and 4 hold, and let $\{x_k, \lambda_k\}$ denote the primal-dual iterate sequence generated by Algorithm 1 when $\{\alpha_k\}$ is chosen as $\alpha_k = \frac{\alpha_0}{(k+1)^{2(1+c)}}$ for $k \geq 0$ for some $\alpha_0 > 0$ and $c > 0$. Then, for all $\epsilon \in (0, 1)$, there exists a $k(\epsilon) \in \mathbb{Z}_+$ such that $|f(\bar{x}_k; \theta^*) - f^*| \leq \epsilon$ and $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \epsilon$ for all $k \geq k(\epsilon) = \mathcal{O}(\epsilon^{-2})$ and requires at most $\mathcal{O}(\epsilon^{-4})$ proximal-gradient computations as shown in Step 1 of Algorithm 2, where $\bar{x}_k \triangleq \frac{1}{k} \sum_{i=1}^k x_i$ for $k \geq 1$.*

Proof. To simplify the notation throughout the proof, let $\gamma \triangleq \sum_{i=0}^{\infty} \sqrt{\alpha_k} < +\infty$ and $\eta \triangleq \kappa \frac{\|\theta_0 - \theta^*\|}{1-\tau}$. According to Proposition 3.2, $\|\lambda_k - \lambda^*\| \leq C_\lambda$ for all $k \geq 1$, where

$$C_\lambda \triangleq \sqrt{2\rho} \gamma + \rho \eta + \|\lambda_0 - \lambda^*\|.$$

In addition, according to Theorem 3.3, $0 \leq f^* - g_\rho(\bar{\lambda}_k; \theta^*) \leq B_g/k$ for all $k \geq 1$, where

$$B_g \triangleq \frac{1}{2\rho} \|\lambda_0 - \lambda^*\|^2 + C_\lambda \left(\sqrt{\frac{2}{\rho}} \gamma + \eta \right) \quad (4.5)$$

and $\bar{\lambda}_k \triangleq \frac{1}{k} \sum_{i=1}^k \lambda_i$ for $k \geq 1$. By choosing $\alpha_0 > 0$ appropriately, we can generate an $\{\alpha_k\}$ sequence such that $\gamma = \frac{1}{\sqrt{2\rho}}$. Hence,

$$C_\lambda = 1 + \rho \eta + \|\lambda_0 - \lambda^*\| \text{ and } B_g = \frac{1}{2\rho} \|\lambda_0 - \lambda^*\|^2 + C_\lambda \left(\frac{1}{\rho} + \eta \right). \quad (4.6)$$

Moreover, according to Theorem 3.6, $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \mathcal{V}(k)$ for all $k \geq 1$, where

$$\mathcal{V}(k) = \frac{C_1}{\sqrt{k}} + \frac{C_2}{k}, \quad C_1 \triangleq \sqrt{\frac{2B_g}{\rho} + \left(\frac{C_\lambda}{\rho}\right)^2}, \quad C_2 \triangleq \frac{1}{\rho} + \mu,$$

and $\mu \triangleq \frac{(L_h + \kappa)\|\theta_0 - \theta^*\|}{1-\tau}$. Using (4.6), we expand C_1^2 as follows:

$$\begin{aligned} C_1^2 &= \frac{1}{\rho^2} \|\lambda_0 - \lambda^*\|^2 + \frac{2}{\rho} \left(\frac{1}{\rho} + \eta \right) C_\lambda + \left(\frac{C_\lambda}{\rho} \right)^2 \\ &= \frac{1}{\rho^2} \|\lambda_0 - \lambda^*\|^2 + \frac{2}{\rho} \left(\frac{1}{\rho} + \eta \right) (1 + \rho \eta + \|\lambda_0 - \lambda^*\|) + \frac{1}{\rho^2} (1 + \rho \eta + \|\lambda_0 - \lambda^*\|)^2 \\ &= \frac{1}{\rho^2} \left[2(1 + \rho \eta + \|\lambda_0 - \lambda^*\|)^2 + (1 + \rho \eta)^2 \right]. \end{aligned}$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have

$$C_1 \leq (\sqrt{2} + 1) \left(\frac{1}{\rho} + \eta \right) + \frac{\sqrt{2}}{\rho} \|\lambda_0 - \lambda^*\|.$$

Hence, we can upper bound $\mathcal{V}(k)$ in terms of problem parameters for all $k \geq 1$:

$$\mathcal{V}(k) \leq \frac{1}{\sqrt{k}} \left(\frac{1}{\rho} \left(\sqrt{2} + 1 + \sqrt{2} \|\lambda_0 - \lambda^*\| \right) + \eta(\sqrt{2} + 1) \right) + \frac{1}{k} \left(\frac{1}{\rho} + \mu \right) \quad (4.7)$$

$$= \frac{1}{\rho} \left(\frac{1}{k} + \frac{\sqrt{2}(\|\lambda_0 - \lambda^*\| + 1) + 1}{\sqrt{k}} \right) + \frac{\mu}{k} + \frac{(\sqrt{2} + 1) \eta}{\sqrt{k}}. \quad (4.8)$$

To further simplify the notation, let $\Gamma_\theta \triangleq (\sqrt{2} + 1)\eta$ and $\Gamma_\lambda \triangleq \sqrt{2}(\|\lambda_0 - \lambda^*\| + 1) + 1$. Next, given $\epsilon \in (0, 1)$, let

$$\rho = \rho_o \epsilon^{-r}, \quad k(\epsilon, s) \triangleq \lceil k_o \epsilon^{-s} \rceil$$

for some fixed $r, s \geq 0$ and $\rho_o, k_o > 0$, where ρ_o and k_o depend only on problem parameters and are independent of solution accuracy ϵ . From (4.8), it immediately follows that for all $k \geq k(\epsilon, s)$

$$\begin{aligned} \mathcal{V}(k) &\leq \frac{1}{\rho_o \epsilon^{-r}} \left(\frac{1}{k_o \epsilon^{-s}} + \frac{\Gamma_\lambda}{\sqrt{k_o \epsilon^{-s}} \sqrt{k_o \epsilon^{-s}}} \right) + \frac{\mu}{k_o \epsilon^{-s}} + \frac{\Gamma_\theta}{\sqrt{k_o \epsilon^{-\frac{s}{2}}}} \\ &= \frac{1}{\rho_o k_o} \epsilon^{r+s} + \frac{\Gamma_\lambda}{\rho_o \sqrt{k_o}} \epsilon^{r+\frac{s}{2}} + \frac{\mu}{k_o} \epsilon^s + \frac{\Gamma_\theta}{\sqrt{k_o}} \epsilon^{\frac{s}{2}} \triangleq \bar{\mathcal{V}}(\epsilon, r, s, \rho_o, k_o). \end{aligned} \quad (4.9)$$

Note that both $\epsilon^r < 1$ and $\epsilon^{s/2} < 1$ for any $r, s \geq 0$ since $\epsilon \in (0, 1)$.

Since $\mathcal{V}(k)$ decreases with $\mathcal{O}(1/\sqrt{k})$ rate, the suboptimality bounds obtained in Theorem 3.7 satisfy $\frac{\rho}{2} \mathcal{V}^2(k) + \|\lambda^*\| \mathcal{V}(k) \geq U/k$ for all $k \geq \lceil k_o \epsilon^{-s} \rceil$ when $\epsilon \in (0, 1)$ is sufficiently small. Although the proof can be written for all $\epsilon \in (0, 1)$, we assume that $\epsilon > 0$ is sufficiently small to simplify the notation in the rest of this section; therefore, $|f(\bar{x}_k; \theta^*) - f^*| \leq \epsilon$ and $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \epsilon$ for all $k \geq k(\epsilon, s)$ whenever k also satisfies

$$\frac{\rho}{2} \mathcal{V}^2(k) + \|\lambda^*\| \mathcal{V}(k) \leq \epsilon, \quad \mathcal{V}(k) \leq \epsilon. \quad (4.10)$$

Clearly, (4.10) holds whenever $\frac{\rho}{2} \mathcal{V}^2(k) \leq \frac{\epsilon}{2}$ and $(\|\lambda^*\| + 1) \mathcal{V}(k) \leq \frac{\epsilon}{2}$; hence, using (4.9), we can conclude that given sufficiently small $\epsilon > 0$, \bar{x}_k is ϵ -optimal and ϵ -feasible for all $k \geq k(\epsilon, s)$ if $r, s \geq 0$ and $\rho_o, k_o > 0$ satisfy the following sufficient condition

$$\bar{\mathcal{V}}(\epsilon, r, s, \rho_o, k_o) \leq \min \left\{ \frac{\epsilon^{\frac{r+1}{2}}}{\sqrt{\rho_o}}, \frac{\epsilon}{2\|\lambda^*\| + 2} \right\}. \quad (4.11)$$

Next, among all $r, s \geq 0$ and $\rho_o, k_o > 0$ satisfying this sufficient condition, we investigate the optimal choice that minimizes the overall computational complexity corresponding to $k(\epsilon, s)$ outer iterations to achieve ϵ -accuracy. According to Step 1 in Algorithm 1, we need to ensure α_k level accuracy in function values at iteration k of the outer loop; hence, according to Lemma 4.2, starting from x_k , we need to perform at most $\sqrt{\frac{2}{\alpha_k} L_{\nu, x}(\rho, \theta_k)} \|x_k - x_{k+1}^*\|$ inner iterations, where x_{k+1}^* is an arbitrary optimal solution to (4.4). Recalling that $\|x - x'\| \leq 2D_x$ for any $x, x' \in X$ (see Remark 2.1), the number of inner iterations within the k -th outer loop can be bounded as follows for $k \geq 0$,

$$\begin{aligned} \sqrt{\frac{2L_{\nu, x}(\rho, \theta_k)}{\alpha_k}} \|x_k - x_{k+1}^*\| &\leq \sqrt{\frac{8(L_{p, x} + \rho L_{h, x}^2)}{\frac{\alpha_0}{(k+1)^{2(1+c)}}}} D_x \\ &= \sqrt{\frac{8(L_{p, x} + \rho L_{h, x}^2)}{\alpha_0}} D_x (k+1)^{1+c}. \end{aligned}$$

Recall that $\alpha_0 > 0$ is chosen such that $\gamma = \sum_{k=0}^{\infty} \sqrt{\alpha_k} = \frac{1}{\sqrt{2\rho}}$. Hence, we have

$$\begin{aligned} \frac{1}{\sqrt{2\rho}} &= \sum_{k=0}^{\infty} \sqrt{\alpha_k} = \sqrt{\alpha_0} \sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^{1+c} \\ &\leq \sqrt{\alpha_0} \left[1 + \int_{t=1}^{\infty} \left(\frac{1}{t}\right)^{1+c} dt \right] = \sqrt{\alpha_0} \left(1 + \frac{1}{c} \right). \end{aligned}$$

Hence, for $c \in (0, 1)$, we get $\frac{1}{\alpha_0} \leq \frac{8\rho}{c^2}$. Moreover, we also have

$$\sum_{k=0}^{k(\epsilon, s)} (k+1)^{1+c} \leq \int_0^{k(\epsilon, s)} (t+1)^{1+c} dt.$$

Therefore, the total number of inner iterations to obtain ϵ accuracy is bounded by

$$\begin{aligned} &\frac{8D_x}{c} \sqrt{\rho (L_{p,x} + \rho L_{h,x}^2)} \sum_{k=0}^{k(\epsilon, s)} (k+1)^{1+c} \\ &\leq \frac{8D_x}{c(2+c)} \sqrt{\rho_o L_{p,x} + (\rho_o L_{h,x})^2 \epsilon^{-r}} ([k_o \epsilon^{-s}] + 1)^{2+c} \epsilon^{-r/2}, \end{aligned} \quad (4.12)$$

where $\rho = \rho_o \epsilon^{-r}$ and $k(\epsilon, s) = \lceil k_o \epsilon^{-s} \rceil$. According to (4.9), for (4.11) to hold for all sufficiently small ϵ , we require $s \geq \max\{r+1, 2\}$. Hence, the best achievable rate is obtained when $r = 0$ and $s = 2$, which results in $\tilde{\mathcal{O}}(\epsilon^{-4})$ rate. Indeed, choosing $r = 0$, $s = 2$, $\rho_o \geq 4(\|\lambda^*\| + 1)^2$ and k_o such that $\sqrt{\rho_o} k_o - (\Gamma_\lambda + \Gamma_\theta \rho_o) \sqrt{k_o} - (1 + \mu \rho_o) \geq 0$ satisfies the sufficient condition in (4.11). \square

We conclude with a corollary that specifies the iteration complexity associated with the perfectly specified problem in which $\theta_0 = \theta^*$; hence, Assumption 2 implies that $\theta_k = \theta^*$ for all $k \geq 0$.

COROLLARY 4.4. *Suppose $\theta_0 = \theta^*$ in Algorithm 1. Under the same assumptions as stated in Theorem 4.3, for all sufficiently small $\epsilon > 0$, there exist $\rho = \mathcal{O}(1/\epsilon)$ and $k_o \in \mathbb{Z}_+$, independent of $\epsilon > 0$, such that $|f(\bar{x}_k; \theta^*) - f^*| \leq \epsilon$ and $d_{-\mathcal{K}}(h(\bar{x}_k; \theta^*)) \leq \epsilon$ for all $k \geq k_o$ which requires at most $\mathcal{O}(\epsilon^{-1})$ proximal-gradient computations as shown in Step 1 of Algorithm 2, where $\bar{x}_k \triangleq \frac{1}{k} \sum_{i=1}^k x_i$ for $k \geq 0$.*

Proof. When $\theta_0 = \theta^*$, we have $\Gamma_\theta = \mu = 0$. Moreover, according to (4.9) and (4.11), $s \geq 2(1-r)$; hence, we obtain the best achievable rate for the overall iteration complexity shown in (4.12) when $r = 1$ and $s = 0$, which results in $\mathcal{O}(\epsilon^{-1})$ rate. Indeed, choosing $r = 1$, $s = 0$, $\rho_o \geq 4(\|\lambda^*\| + 1)^2$ and k_o such that $\sqrt{\rho_o} k_o - \Gamma_\lambda \sqrt{k_o} - 1 \geq 0$ satisfies the new sufficient condition in (4.11). \square

Remark 4.1. *Suppose we set $\lambda_0 = \mathbf{0}$. The proof of Corollary 4.4 shows that for $\rho_o \geq 4(\|\lambda^*\| + 2)^2$, $k_o = 1$ satisfies the sufficient condition, $\sqrt{\rho_o} k_o - \Gamma_\lambda \sqrt{k_o} - 1 \geq 0$. Therefore, setting $\rho = \rho_o/\epsilon$ and computing one outer iteration is sufficient. Indeed, according to (4.12), implementing APG on $\min_{x \in X} \mathcal{L}_\rho(x, \lambda_0; \theta^*) = f(x; \theta^*) + \frac{\rho}{2} d_{-\mathcal{K}}^2(h(x; \theta^*))$ will generate an ϵ -optimal and ϵ -feasible solution to the original problem $\mathcal{C}(\theta^*)$ within $\mathcal{O}(D_x \rho_o L_{h,x} \frac{1}{\epsilon})$ iterations.*

4.2. Overall iteration complexity for increasing $\{\rho_k\}$. As shown in Theorem 3.11, when $\rho_k = \rho_0 \beta^k$ for some $\beta > 1$ and $\rho_0 > 0$, we obtain a geometric rate of convergence of sub-optimality error in terms of outer iterations of Algorithm 1 in

the form of B_k/β^k such that $B_{k+1} < B_k$ for all k ; hence, $\sup_k B_k < +\infty$. While this may seem to be promising at first glance, it should be noted that as k increases, ρ_k also increases geometrically; hence, $L_{\nu,x}(\rho_k, \theta_k)$, the Lipschitz constant of $\nabla_x \nu_{\rho_k}(\cdot; \theta_k)$ increases at a geometric rate as well (see Lemma 4.1), which adversely affects the convergence rate of APG (see Lemma 4.2). Therefore, increasing $\{\rho_k\}$ has two distinct effects: on one side, compared to constant ρ , increasing $\{\rho_k\}$ increases the rate of convergence of outer iteration from $\mathcal{O}(\frac{1}{\sqrt{k}})$ to $\mathcal{O}(\frac{1}{\beta^k})$; while on the other hand, it also increases the complexity of the inner computation. The following theorem derives the overall iteration complexity of Algorithm 1 for the increasing penalty sequence.

THEOREM 4.5. *Let Assumptions 1, 2 and 4 hold. Let $\{x_k, \lambda_k\}$ be the primal-dual iterate sequence generated by Algorithm 1 for the parameter sequences $\{\alpha_k\}_{k \in \mathbb{Z}}$ and $\{\rho_k\}_{k \in \mathbb{Z}}$ as defined in Theorem 3.11. Then, for all $\epsilon > 0$, there exists a $k(\epsilon) \in \mathbb{Z}_+$ such that $|f(x_k; \theta^*) - f^*| \leq \epsilon$ and $d_{-\mathcal{K}}(h(x_k; \theta^*)) \leq \epsilon$ for all $k \geq k(\epsilon) = \mathcal{O}(\log(\epsilon^{-1}))$ which requires at most $\mathcal{O}(\epsilon^{-1} \log(\epsilon^{-1}))$ proximal-gradient computations as shown in Step 1 of Algorithm 2.*

Proof. Suppose $\{\alpha_k\}_{k \in \mathbb{Z}}$ and $\{\rho_k\}_{k \in \mathbb{Z}}$ chosen as in Theorem 3.11. (3.38) and (3.40) together with the inequality $(2C'_\lambda + \max\{\|\lambda^*\|, 1\})^2 \geq 2C'_\lambda$ clearly imply that for all $k \geq 0$

$$\begin{aligned} & \max\{|f(x_{k+1}; \theta^*) - f^*|, d_{-\mathcal{K}}(h(x_{k+1}; \theta^*))\} \\ & \leq \frac{1}{\rho_0} (2C'_\lambda + \max\{\|\lambda^*\|, 1\})^2 \\ & \quad + \rho_0 \left(L_{h,\theta} \|\theta_0 - \theta^*\| \delta^k + \frac{\max\{L_f, L_{h,\theta}\}}{\rho_0 L_{h,\theta}} \right)^2 + \frac{\alpha_0}{(k+1)^{2(1+c)}} \triangleq \bar{B}_k. \end{aligned}$$

Moreover, $0 < \bar{B}_k \leq \bar{B}_0$ for all $k \geq 0$. Therefore, for any given $\epsilon > 0$, Theorem 3.11 implies that it takes at most $k(\epsilon) = \log_\beta(\frac{\bar{B}_0}{\epsilon})$ outer iterations to achieve ϵ -optimal and ϵ -feasible primal solution. According to Step 1 in Algorithm 1 we need to ensure α_k level accuracy in function values at iteration k of the outer loop; hence, according to Lemma 4.2, starting from x_k , we need to perform $\sqrt{\frac{2}{\alpha_k} L_{\nu,x}(\rho_k, \theta_k)} \|x_k - x_{k+1}^*\|$ inner iterations, where x_{k+1}^* is an arbitrary optimal solution to (4.4). Recalling that $\|x - x'\| \leq 2D_x$ for any $x, x' \in X$ (see Remark 2.1), the number of inner iterations within the k -th outer loop can be bounded as follows for $k \geq 0$,

$$\sqrt{\frac{2}{\alpha_k} L_{\nu,x}(\rho_k, \theta_k)} \|x_k - x_{k+1}^*\| \leq \beta^k (k+1)^{1+c} \underbrace{\sqrt{\frac{2}{\alpha_0} \left(\frac{L_{p,x}}{\beta^k} + \rho_0 L_{h,x}^2 \right)}}_{\triangleq M_k} D_x.$$

Clearly, $0 < M_k \leq M_0$ for all $k \geq 0$; hence, the overall number of inner iterations to obtain ϵ accuracy can be bounded above as follows:

$$\begin{aligned} \sum_{k=0}^{\log_\beta(\frac{\bar{B}_0}{\epsilon})} M_0 \beta^k (k+1)^{1+c} & \leq M_0 \left(\log_\beta \left(\frac{\bar{B}_0}{\epsilon} \right) + 1 \right)^{(1+c)} \sum_{k=0}^{\log_\beta(\frac{\bar{B}_0}{\epsilon})} \beta^k \\ & \leq M_0 \left(\log_\beta \left(\frac{\bar{B}_0}{\epsilon} \right) + 1 \right)^{1+c} \frac{\beta}{\beta-1} \left(\frac{\bar{B}_0}{\epsilon} \right). \end{aligned}$$

□

In following remark, we note that when θ^* is known, i.e., when $\theta_0 = \theta^*$, then the order of overall iteration complexity of Algorithm 1 remains unchanged, i.e., $\tilde{\mathcal{O}}(\epsilon^{-1} \log(\epsilon^{-1}))$ as in the case when learning is involved. However, there is a reduction in the $\mathcal{O}(1)$ constant.

Remark 4.2. *Under the same assumptions as stated in Theorem 4.5, when $\theta_0 = \theta^*$ in Algorithm 1, then the bounds in (3.38) can be modified as follows:*

$$\begin{aligned} |f(x_{k+1}, \theta^*) - f^*| &\leq \frac{1}{\beta^k} \left[\frac{1}{\rho_0} (2C'_\lambda + \|\lambda^*\|)^2 + \alpha_0 \right], \\ \text{and } d_{-\mathcal{K}}(h(x_{k+1}; \theta^*)) &\leq \frac{1}{\beta^k} \frac{2C'_\lambda}{\rho_0}; \end{aligned}$$

therefore, as in Theorem 4.5, the overall number of inner iterations to obtain ϵ accuracy can be bounded above by

$$\sum_{k=0}^{\log_\beta \left(\frac{B'}{\epsilon} \right)} M' \beta^k (k+1)^{1+c} \leq M' \left(\log_\beta \left(\frac{B'}{\epsilon} \right) + 1 \right)^{1+c} \frac{\beta}{\beta-1} \left(\frac{B'}{\epsilon} \right) = \tilde{\mathcal{O}}(\epsilon^{-1} \log(\epsilon^{-1})),$$

where $M' = \sqrt{\frac{2}{\alpha_0} (L_{p,x} + \rho_0 \|A(\theta^*)\|^2)}$ and $B' = \frac{1}{\rho_0} (2C'_\lambda + \max\{\|\lambda^*\|, 1\})^2 + \alpha_0$, in which the constant (see Lemma 3.8) reduces to $C'_\lambda \triangleq \sum_{i=0}^\infty \sqrt{2\rho_i \alpha_i} + \|\lambda_0 - \lambda^*\|$.

5. Numerical Results. In this section, we present a problem related to portfolio optimization in Section 5.1 and define the problem parameters in Section 5.2. The empirical performance of the proposed algorithms is examined in Section 5.3.

5.1. Problem description. In this subsection, we first describe the computational problem and then the learning problem.

Computational Problem: We consider the Markowitz portfolio optimization problem [20], where $\{\mathcal{R}_i\}_{i=1}^n$ denote the *random* returns for n financial assets. Assume that the joint distribution of aggregated return (as given by $\mathcal{R} = [\mathcal{R}_i]_{i=1}^n$) is a multivariate normal distribution with mean vector $\mathbb{R}^n \ni \mu^o := \mathbb{E}[\mathcal{R}]$ and covariance matrix $\mathbb{R}^{n \times n} \ni \Sigma^o := \mathbb{E}[(\mathcal{R} - \mu^o)^\top (\mathcal{R} - \mu^o)]$. We assume that $\Sigma^o = [\sigma_{ij}]_{1 \leq i, j \leq n}$ is *positive definite*, implying that there are no redundant assets in our collection. Suppose $x_i \in \mathbb{R}$ denotes the proportion of asset i in the portfolio held throughout the given period. Hence, $x = [x_i]_{i=1}^n \in \mathbb{R}^n$ such that $\sum_{i=1}^n x_i = 1$ and $x_i \geq 0$ for all $i = 1, \dots, n$ corresponds to a portfolio with *no short selling*. Practitioners often use additional constraints to reduce *sector risk* by grouping together investments in securities of a sector and setting a limit on the exposure to this sector [7]. Suppose there are s sectors, and m_j is the maximum proportion of the portfolio that can be invested in sector j for $j = 1, \dots, s$. Let $I_j \subset \{1, \dots, n\}$ be the set of indices corresponding to the assets belonging to sector j for $j = 1, \dots, s$. Note that the same asset can belong to more than one sectors; hence, we do *not* assume that $\{I_j\}_{j=1}^s$ is a partition. These sector constraints can be written as

$$\sum_{i \in I_j} x_i \leq m_j, \text{ for } j = 1, \dots, s.$$

Clearly, the above set of constraints can be represented by the matrix notation $Ax \leq b$, where $b \triangleq [m_1, \dots, m_s]^\top$ and $A \in \mathbb{R}^{s \times n}$ such that $A_{ji} = 1$ if asset i belongs to sector j , and is 0 otherwise. To decide the optimal portfolio, we face two competing

objectives: minimize the risk, i.e., the variance of the portfolio return, and maximize the expected portfolio return. This portfolio selection model is called the Markowitz portfolio optimization problem; given a trade-off parameter $\kappa > 0$, some estimates μ and Σ for μ^o and Σ^o , resp., it can be written as follows:

$$(\mathcal{C}(\Sigma)) : \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} x^T \Sigma x - \kappa \mu^T x : Ax \leq b, x \in X \right\}, \quad (5.1)$$

where $X \triangleq \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x \geq 0\}$. A lower value of κ leads to a “risk averse” portfolio while larger values correspond to “risk seeking” ones. To compute an optimal solution of $\mathcal{C}(\Sigma)$, one may use constrained optimization techniques when Euclidean projections onto the polyhedral set, defined by $Ax \leq b$ and $x \in X$, cannot be computed efficiently. Note that when $\{I_j\}_{j=1}^s$ is not a partition, i.e., $I_j \cap I_k \neq \emptyset$ for some $1 \leq j \neq k \leq s$, then it may usually be not efficient to compute projections at each iteration. Hence, one may overcome the projection requirement by relaxing the hard constraint, $Ax \leq b$, and adopting an augmented Lagrangian scheme.

If the *true* values of the parameters μ^o and Σ^o are known, then the Markowitz problem is just a convex quadratic optimization problem over a polyhedral set. However, knowing the true values of μ^o and Σ^o often cannot be taken for granted. In fact, even the estimation of these parameters is generally not easy. In this section, we consider a setting where true μ^o vector is specified, i.e., $\mu = \mu^o$ but the true covariance matrix Σ^o is *unknown*; but, it can be computed as the optimal solution to a suitably defined *learning problem*.

Learning Problem: Given a sample of returns for n assets such that sample size is equal to p for each asset, let $S = (s_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ denote the sample covariance matrix. In practice, we usually have $p \ll n$, which means that the number of assets is far greater than the sample size. Since $p < n$, S cannot be positive definite; on the other hand, $\Sigma^o \succ \mathbf{0} \in \mathbb{R}^{n \times n}$. Hence, instead of using S as our true covariance estimator, we consider the sparse covariance selection (SCS) problem, proposed in [31], as our learning problem and is defined below:

$$\Sigma^* := \operatorname{argmin}_{\Sigma \in S^n} \left\{ \frac{1}{2} \|\Sigma - S\|_F^2 + v |\Sigma|_1 : \Sigma \succeq \epsilon I \right\},$$

where v and ϵ are positive regularization parameters, S^n denotes the set of $n \times n$ symmetric matrices, $\|\cdot\|_F$ is the Frobenius norm, $|\cdot|_1$ is the ℓ_1 norm of all off-diagonal elements treated as a column vector, and $\Sigma \succeq \epsilon I$ implies that all eigenvalues of Σ are greater than equal to $\epsilon > 0$. Notice that the constraint in this problem guarantees that the estimate Σ^* is positive definite, and the ℓ_1 regularization term in the objective promotes sparsity in Σ^* . Therefore, the optimal solution Σ^* will satisfy our full-rank assumption on the covariance matrix. Lack of such properties may lead to undesirable under-estimation of risk in the high dimensional Markowitz problem and also may cause the corresponding optimization problem to be ill defined [31]. Hence, we assume that Σ^* can be safely used to approximate Σ^o .

We choose to solve the SCS learning problem using an ADMM algorithm. In order to apply this scheme, we adopt a variant of the formulation presented in [31]: we first introduce a new variable Φ and an equality constraint as follows:

$$\min_{\Sigma, \Phi \in S^n} \{1_Q(\Sigma) + \frac{1}{2} \|\Sigma - S\|_F^2 + v |\Phi|_1 : \Sigma = \Phi\} \quad (\text{SCS})$$

where 1_Q denotes the indicator function of the set $Q \triangleq \{\Sigma \in S^n : \Sigma \succeq \epsilon I\}$. Now, (SCS) is in a form where we may easily apply the ADMM scheme. Let $\gamma : S^n \rightarrow \mathbb{R}$

such that $\text{dom}(\gamma) = Q$, and $\gamma(\Sigma) = \frac{1}{2}\|\Sigma - S\|_F^2$. Since γ is strongly convex with a Lipschitz continuous gradient over its domain, then the ADMM algorithm generates a sequence $\{\Sigma_k\}$, guaranteed to converge at a linear rate to the optimal solution Σ^* [9], i.e., $\|\Sigma_k - \Sigma^*\| \leq \tau^k \|\Sigma_0 - \Sigma^*\|$, for some $\tau \in (0, 1)$. Hence, Assumption 2 is satisfied.

5.2. Problem parameters. Suppose there are $n = 1500$ available assets from $s = 10$ sectors. To construct the learning problem, we first generate the true parameters, namely the mean return μ^o and covariance matrix $\Sigma^o = [\sigma_{ij}]_{1 \leq i, j \leq n}$ based on the following rules: μ^o is generated from a uniform distribution over the hyperbox $[-1, 1]^n \in \mathbb{R}^n$ while $\sigma_{ij} = \max\{1 - (|i - j|/10), 0\}$. Next, we generate p i.i.d. samples of random returns $\{\mathcal{R}_t\}_{t=1}^p$ from a multivariate normal distribution with mean μ^o and covariance matrix Σ^o where the sample size is set to $p = \frac{n}{2}$. Sample returns are then used to calculate the sample covariance matrix S . Given that μ^0 is known, Σ^* is the solution to the learning problem (SCS) with $v = 0.4$. Consequently, the optimal portfolio, $x^* \in X$, is the solution to $\mathcal{C}(\Sigma^*)$, where $\kappa = 0.1$.

5.3. Empirical analysis of performance. In this subsection, we conduct empirical studies of the constant penalty parameter and increasing penalty parameter schemes. We conclude with a brief discussion on how the proposed simultaneous schemes compare with their sequential counterparts. In all tables, CPU times are reported in *seconds*.

5.3.1. Constant penalty parameter ρ . In the first set of experiments, we assume that Σ^* is known, implying that the problem is perfectly specified, and investigate the performance of the inexact augmented Lagrangian scheme proposed in Algorithm 1 using the sequence of inexactness $\{\alpha_k\}$ stated in Theorem 4.3. Recall that, given ϵ , Theorem 4.3 and Corollary 4.4 suggest to use the following penalty parameter ρ and sequence of inexactness $\{\alpha_k\}$ in order to obtain the best overall iteration complexity of $O(\frac{1}{\epsilon})$: $\rho = \frac{\rho_0}{\epsilon}$ and $\alpha_k = \alpha_0 k^{-2(1+c)}$, where α_0 is chosen such that $\sum_{k=0}^{\infty} \sqrt{\alpha_k} = \frac{1}{\sqrt{2\rho}}$. We choose $\rho_0 = 1$ and $c = 1e-3$. Note that starting from x_{k-1} , Algorithm 2 (APG) is employed to find an α_k -optimal solution x_k in the inner loop step. Lemma 4.2 states the required number of APG iterations to obtain such an x_k . As stated in this lemma, that number is bounded above by $\sqrt{2L_{\nu,x}/\alpha_k} \|x_{k-1} - x_k^*\|$, where $L_{\nu,x}$ denotes the Lipschitz constant of $\nabla_x \mathcal{L}_\rho(x, \lambda_k; \Sigma_k)$ and $x_k^* = \text{argmin} \mathcal{L}_\rho(x, \lambda_k; \Sigma_k)$. Note that $L_{\nu,x} = \sigma_{\max}(\Sigma_k) + \rho \sigma_{\max}^2(A)$.

Table 5.1 details the sub-optimality, infeasibility, and the associated computational effort to obtain an ϵ -optimal and ϵ -feasible solution for various values of ϵ when Σ^* is available. Specifically, we list the the number of outer and inner iterations, where the number of inner iterations is that required by utilizing the rate statement for the APG scheme. We now compare the results from Table 5.1 with

TABLE 5.1
Solution quality and computational effort: Constant ρ and known Σ^* .

ϵ	$\frac{ f(\bar{x}_K, \Sigma^*) - f^* }{ f^* }$	$d_{\mathbb{R}^m}(\bar{A}\bar{x}_K - b)$	# outer (K)	# inner	CPU time
1e-1	8.2e-2	1.1e-3	4	23	12
1e-2	7.3e-3	3.4e-4	5	63	34
1e-3	7.4e-4	4.5e-4	4	75	41
1e-4	9.7e-5	7.0e-5	4	134	69

those obtained by implementing Algorithm 1 using the learning sequence $\{\Sigma_k\}$ for misspecified parameter Σ^* . According to Theorem 4.3, to obtain best overall iteration complexity of $\mathcal{O}(\epsilon^{-4})$, we have to choose α_k and ρ such that $\rho = \rho_0 > 0$ and

TABLE 5.2
Solution quality and computational effort: Constant ρ and misspecified Σ^*

ϵ	$\frac{ f(\bar{x}_K; \Sigma^*) - f^* }{ f^* }$	$\frac{\ \Sigma_K - \Sigma^*\ }{\ \Sigma^*\ }$	$d_{\mathbb{R}^m}^m(A\bar{x}_K - b)$	# outer (K)	# inner	CPU time	
						learn	opt.
1e-1	8.6e-2	1.1e-1	1.2e-3	4	15	100	7
1e-2	8.8e-3	1.9e-1	1.2e-4	5	65	125	32
1e-3	9.9e-4	1.7e-2	3.7e-5	16	1383	400	692
1e-4	3.1e-4	6.9e-3	2.6e-6	36	9588	900	4733

$\alpha_k = \alpha_0/k^{2(1+c)}$, where α_0 satisfies $\sum_{k=0}^{\infty} \sqrt{\alpha_k} = 1/\sqrt{2\rho}$. We choose $\rho_0 = 1$ and $c = 1e-3$. Table 5.2 lists the results for various values of ϵ . In addition, we compare the CPU time spent for computation versus learning. Note that while our theoretical bound requires at least $\mathcal{O}(\epsilon^{-4})$ overall number of iterations, the empirical behavior is far better, suggesting that the bound obtained in Theorem 4.3 is not tight and will require further study. When comparing Tables 5.1 and 5.2, we note that the overall effort in the misspecified regime is significantly larger. This is not surprising, since Table 5.1 does not include the effort to provide an exact Σ^* . The comparison of running times provided in Table 5.2 suggests that the effort for learning is by no means modest.

In Figure 5.1(left), we provide a graphical representation of how the empirically observed primal suboptimality error changes with K , the number of outer iterations when $\epsilon = 1e-2$. This graph is overlaid by the theoretical bound based on Theorem 3.7. In addition, Figure 5.1(right) displays the corresponding primal infeasibility and the associated theoretical bound obtained in Theorem 3.6.

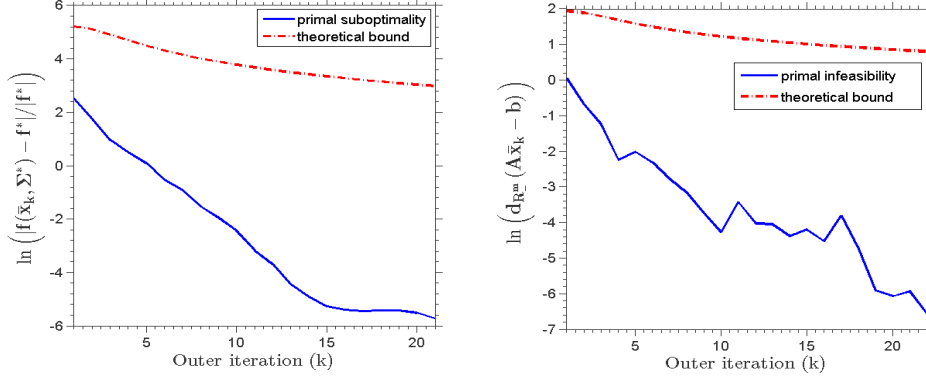


FIG. 5.1. Empirical error vs theoretical bound using constant ρ : (left) Primal suboptimality $\ln(|f(\bar{x}_k; \Sigma^*) - f^*|/|f^*|)$; and (right) Primal infeasibility $\ln(d_{\mathbb{R}^m}^m(A\bar{x}_k - b))$.

5.3.2. Increasing penalty parameter sequence $\{\rho_k\}$. Next, we examine the computational performance of Algorithm 1 for an increasing sequence of penalty parameters, $\{\rho_k\}$. To achieve the overall iteration complexity of $\mathcal{O}(\epsilon^{-1} \log(\epsilon^{-1}))$, Theorem 3.11 suggests using sequences ρ_k and α_k where $\rho_k = \rho_0 \beta^k$ and $\alpha_k = k^{-2(1+c)} \beta^{-k}$, where $\beta \tau < 1$ and τ is such that $\|\Sigma_k - \Sigma^*\| \leq \tau^k \|\Sigma_0 - \Sigma^*\|$. We set $\beta = 1.05$, based on the calculated value of $\tau = 0.91$ and also choose $c = 1e-3$. For various values of ϵ , Tables 5.3 and 5.3.2 display the numerical results for known and misspecified Σ^* , respectively. We begin by noting that the overall complexity in terms of inner iterations is not significantly larger in terms of the number of computational steps as that

observed with known Σ^* , providing empirical support for the theoretical findings of Theorem 4.5 and Remark 4.2. Figure 5.2(left) depicts how the empirically observed primal suboptimality error changes with K when $\epsilon = 1\text{e-}2$. By overlaying the trajectory derived from the non-asymptotic upper bound which diminishes at a linear rate as derived in Theorem 3.11, it is seen that the numerics support the theoretical findings. In addition, Figure 5.2(right) displays the corresponding primal infeasibility and the associated theoretical bound obtained in Theorem 3.11.

Note that when Σ^* is known, as shown in Corollary 4.4, choosing a constant ρ results in $\mathcal{O}(1/\epsilon)$ overall complexity. While in theory, this is preferable to employing an increasing sequence $\{\rho_k\}$ which has a larger complexity of $\mathcal{O}(\epsilon^{-1} \log(\epsilon^{-1}))$, the constant ρ version requires careful estimation of problem parameters and ρ based on the choice ϵ . In contrast, when ρ_k is an increasing sequence, the choice of β and ρ_0 is independent of problem parameters, a distinct advantage of the increasing penalty parameter scheme.

We recall that at iteration k , given α_k, ρ_k, x_k , and θ_k , Algorithm 2 produces an iterate x_{k+1} by proceeding through T_k iterations where T_k is defined in Algorithm 2 and redefined below based on the definition of $L_{\nu,x}(\rho_k, \theta_k)$ and α_k .

$$T_k \triangleq \sqrt{\frac{8(L_{p,x} + \rho_k L_{h,x}^2)}{\alpha_0}} D_x(k+1)^{1+c}. \quad (5.2)$$

In Table 5.3.2, the column “inner (theor.)” refers to the aggregate number of inner iterations, bounded from below by $\sum_k T_k$ where We compare the metric “inner (theor)” with the metric “inner (actual)” which aggregates the number of inner steps of Algorithm 2 to satisfy the error criterion:

$$\mathcal{L}_{\rho_k}(x_{k+1}; \lambda_k, \theta_k) \leq g_{\rho_k}(\lambda_k, \theta_k) + \alpha_k, \quad (5.3)$$

where $\mathcal{L}_{\rho_k}(x; \lambda, \theta)$ is defined in (4.1). This requires computing x_{k+1}^* separately for each iteration and terminating the inner loop when (5.3) holds. Naturally, this is not possible in practice but merely provides a notion of how much lower the complexity of solving the subproblem could be. In fact, the following holds if x_{k+1} denotes the output of Algorithm APG($x_k, \lambda_k, \rho_k, \alpha_k$):

$$[x_{k+1} \leftarrow \mathbf{APG}(x_k, \lambda_k, \alpha_k, \rho_k, T_k)] \implies [x_{k+1} \text{ satisfies (5.3)}].$$

TABLE 5.3
Solution quality and performance statistics: Increasing ρ_k and known Σ^*

ϵ	$\frac{ f(x_K; \Sigma^*) - f^* }{ f^* }$	$d_{\Sigma^*}^m(Ax_K - b)$	# outer (K)	# inner	CPU time
1e-1	7.9e-2	9.2e-3	6	22	13
1e-2	8.0e-3	8.3e-4	12	45	24
1e-3	9.3e-4	1.7e-4	16	92	48
1e-4	2.8e-5	6.3e-5	25	345	166

5.3.3. Sequential vs simultaneous schemes. Our last set of numerics provides a graphical representation of the benefits of simultaneous schemes, and captures the overall effort/time in a single figure (Figure 5.3). To compare our proposed scheme with standard sequential schemes, we incorporate the effort to solve the learning problem in a priori fashion and then use this possibly inexact solution to resolve

TABLE 5.4
Solution quality and performance statistics: Increasing ρ_k and misspecified Σ^*

ϵ	$\frac{ f(x_K; \Sigma^*) - f^* }{ f^* }$	$\frac{\ \Sigma_K - \Sigma^*\ }{\ \Sigma^*\ }$	$d_{\mathbb{R}^m}(A\bar{x}_K - b)$	# outer K	# inner	CPU time	
						learn	opt.
1e-1	9.2e-2	3.4e-1	1.2e-2	5	10	125	4
1e-2	5.3e-3	7.9e-2	1.2e-3	12	58	300	30
1e-3	8.1e-4	4.1e-2	5.5e-4	19	170	474	88
1e-4	9.3e-5	9.7e-3	2.8e-6	48	2600	1200	1330

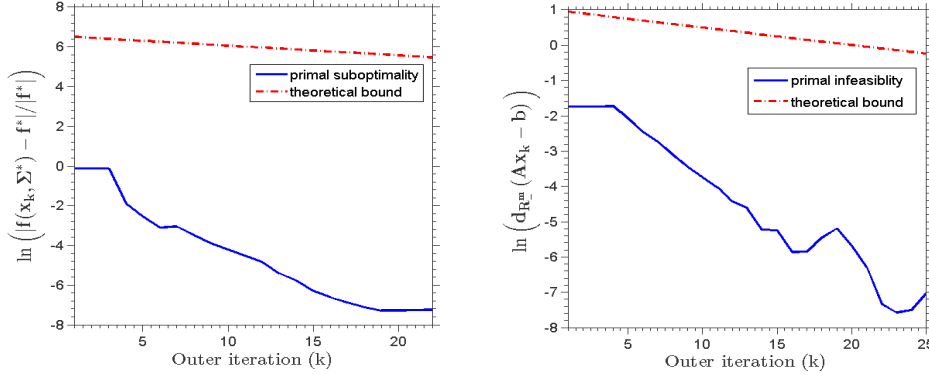


FIG. 5.2. Empirical error vs theoretical bound using increasing ρ_k : (left) Primal relative suboptimality $\ln(|f(x_k; \Sigma^*) - f^*|/|f^*|)$; and (right) Primal infeasibility $\ln(d_{\mathbb{R}^m}(Ax_k - b))$.

the computational problem. For instance, in Figure 5.3, we consider 4 different implementations of the sequential scheme, where the implementations differ by the amount of effort (number of learning steps) employed for obtaining an approximation to Σ^* . On the y -axis, we capture the sub-optimality error and note that while the sequential schemes are making an effort to get an approximation of Σ^* , no improvement is being made in x . Consequently, all of the graphs corresponding to the sequential schemes stay constant. Once an approximation is obtained, the sequential scheme will obtain an approximate solution but the sub-optimality error never diminishes to zero, since the sequential scheme never updates its approximation of Σ^* . The simultaneous approach on the other hand has several benefits: (i) it is characterized by asymptotic convergence, a property that does not hold for sequential schemes; (ii) one can provide non-asymptotic rate bounds for the entire trajectory $\{x_k\}$; and (iii) when it is unclear as to the extent of accuracy required in solving the learning problem, sequential methods can prove to be quite poor while simultaneous schemes perform well.

6. Conclusion. This paper has been motivated by the question of resolving convex optimization problems' plagues due to parametric misspecification both in the objective and the constraint sets. We consider settings where this misspecification may be resolved by solving a suitably defined learning problem and accordingly, we consider the setting where we have two coupled optimization problems; of these, the first one is a misspecified optimization problem where the unknown parameters appear in both the objective function as well as the constraint set, while the second one is a learning problem that arises from having access to a learning data set, collected a-priori. One avenue for contending with such a problem is through an inherently sequential approach that solves the learning problem, and subsequently utilizes this

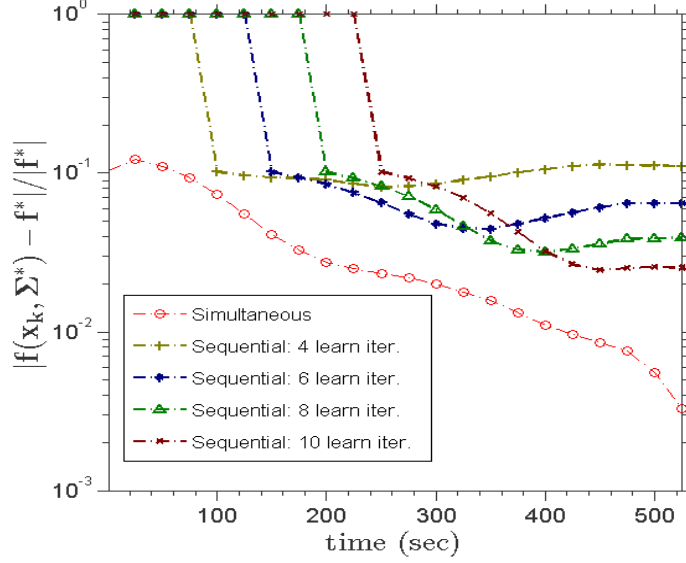


FIG. 5.3. *Simultaneous vs sequential approach – Primal suboptimality for increasing ρ_k : $|f(x_k; \Sigma^*) - f^*|$ in log-scale*

solution in solving the computational problem. Unfortunately, unless accurate solutions of the learning problem are available in a finite number of iterations, sequential approaches may not be advisable due to propagation of error. Instead, we focus on a simultaneous approach that combines learning and computation by adopting inexact augmented Lagrangian (AL) scheme. Two classes of inexact AL schemes have been investigated; first one uses constant penalty parameter in its implementation while second one employs increasing sequence of penalty parameters. In this regard, we make the following contributions: (i) Derivation of the convergence rate for dual optimality, primal infeasibility and primal suboptimality; (ii) Quantification of the learning effect on the rate degradation. (iii) Analysis of overall iteration complexity. Preliminary numerics suggest that the proposed schemes perform well on a misspecified portfolio optimization problem while traditional approaches for addressing misspecification may perform poorly in practice.

REFERENCES

- [1] H. AHMADI, N. S. AYBAT, AND U. V. SHANBHAG, *On the rate analysis of inexact augmented lagrangian schemes for convex optimization problems with misspecified constraints*, in the Proceedings of 2016 American Control Conference (ACC) (forthcoming), 2016. The preprint is made available at <https://arxiv.org/abs/1510.00490> on Oct. 2, 2015.
- [2] H. AHMADI AND U. V. SHANBHAG, *On the resolution of misspecified convex optimization and monotone variational inequality problems*, <http://arxiv.org/abs/1408.5532>.
- [3] —, *Data-driven first-order methods for misspecified convex optimization problems: Global convergence and rate estimates*, in 53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014, IEEE, 2014, pp. 4228–4233.
- [4] NECDET SERHAT AYBAT AND GARUD IYENGAR, *An augmented lagrangian method for conic convex programming*, arXiv preprint arXiv:1302.6322, (2013).
- [5] AMIR BECK AND MARC TEOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences, 2 (2009), pp. 183–202.
- [6] A. BEN-TAL, L. EL GHAOU, AND A. NEMIROVSKI, *Robust optimization*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2009.
- [7] G. CORNUEJOLS AND R. TÜTÜNCÜ, *Optimization Methods in Finance*, Cambridge University Press, 2006. Cambridge Books Online.

- [8] R. W. COTTLE, J-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, Inc., Boston, MA, 1992.
- [9] WEI DENG AND WOTAO YIN, *On the global and linear convergence of the generalized alternating direction method of multipliers*, Journal of Scientific Computing, (2015), pp. 1–28.
- [10] F. FACCHINEI AND J-S. PANG, *Finite-dimensional variational inequalities and complementarity problems. Vol. I*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.
- [11] MAGNUS R. HESTENES, *Multiplier and gradient methods*, Journal of optimization theory and applications, 4 (1969), pp. 303–320.
- [12] JEAN-BAPTISTE HIRIART-URRUTY AND CLAUDE LEMARÉCHAL, *Convex analysis and minimization algorithms II: Advanced theory and bundle methods, vol. 306 of Grundlehren der mathematischen Wissenschaften*, Springer-Verlag, New York, 1993.
- [13] ———, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, Springer-Verlag, New York, 2001.
- [14] H. JIANG AND U. V. SHANBHAG, *On the solution of stochastic optimization problems in imperfect information regimes*, in Winter Simulations Conference: Simulation Making Decisions in a Complex World, WSC 2013, Washington, DC, USA, December 8–11, 2013, IEEE, 2013, pp. 821–832.
- [15] ———, *Data-driven schemes for resolving misspecified mdps: asymptotics and error analysis*, in Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, USA, December 6–9, 2015, IEEE/ACM, 2015, pp. 3801–3812.
- [16] ———, *On the solution of stochastic optimization and variational problems in imperfect information regimes*, SIAM Journal of Optimization (to appear), (2016).
- [17] H. JIANG, U. V. SHANBHAG, AND S. P. MEYN, *Distributed computation of equilibria in misspecified convex stochastic Nash games*, IEEE Transactions on Automatic Control (Conditionally accepted), (2016).
- [18] A. KANNAN, A. NEDIC, AND U. V. SHANBHAG, *Distributed stochastic optimization under imperfect information*, in 54th IEEE Conference on Decision and Control, CDC 2015, Osaka, Japan, December 15–18, 2015, 2015, pp. 400–405.
- [19] GUANGHUI LAN AND RENATO D.C. MONTEIRO, *Iteration-complexity of first-order augmented lagrangian methods for convex programming*, Mathematical Programming, (2015), pp. 1–37.
- [20] HARRY MARKOWITZ, *Portfolio selection**, The Journal of Finance, 7 (1952), pp. 77–91.
- [21] ION NECOARA, ANDREI PATRASCU, AND FRANCOIS GLINEUR, *Complexity certifications of first order inexact lagrangian and penalty methods for conic convex programming*, <http://arxiv.org/abs/1506.05320>.
- [22] Y. NESTEROV, *Gradient methods for minimizing composite functions*, Mathematical Programming, Series B, 140 (2013), pp. 125–161.
- [23] YURI NESTEROV, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.
- [24] M.J.D. POWELL, *Optimization*, Academic Press, London/New York, 1969, ch. A method for nonlinear constraints in minimization problems, p. 283–298.
- [25] R.T. ROCKAFELLAR, *The multiplier method of hestenes and powell applied to convex programming*, Journal of Optimization Theory and Applications, 12 (1973), pp. 555–562.
- [26] R. TYRRELL ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Mathematical Programming, 5 (1973), pp. 354–373.
- [27] R. TYRRELL ROCKAFELLAR, *Augmented lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of operations research, 1 (1976), pp. 97–116.
- [28] R. T. ROCKAFELLAR, *Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming*, Mathematics of Operations Research, 1 (1976), pp. 97–116.
- [29] R. TYRRELL ROCKAFELLAR, *Monotone Operators and the Proximal Point Algorithm*, SIAM J. on Control and Optimization, 14 (1976).
- [30] P. TSENG, *On accelerated proximal gradient methods for convex-concave optimization*. Preprint available at <http://www.eecs.berkeley.edu/~brecht/eecs227cdocs/tseng.pdf>, 2008.
- [31] LINGZHOU XUE, SHIQIAN MA, AND HUI ZOU, *Positive-definite ℓ_1 -penalized estimation of large covariance matrices*, Journal of the American Statistical Association, 107 (2012), pp. 1480–1491.